

BAB IV

HASIL DAN PEMBAHASAN

4.1 *Data Selection*

Tahap data *selection* merupakan langkah awal dalam proses analisis di mana data yang relevan dipilih dari sumber yang tersedia untuk digunakan dalam penelitian. Pada penelitian ini, data diperoleh dari repositori terbuka StatsBomb melalui GitHub, yang menyediakan data *event* pertandingan sepak bola dalam format JSON. Dari seluruh data yang tersedia, hanya data dengan tipe *event Shot* yang dipilih karena fokus penelitian adalah memprediksi kemungkinan terciptanya gol dari sebuah tembakan. Selain itu, hanya kolom-kolom tertentu yang dipilih, seperti informasi tentang teknik tembakan, bagian tubuh yang digunakan, pola permainan, serta posisi awal tembakan, karena kolom-kolom tersebut dianggap memiliki relevansi langsung terhadap peluang mencetak gol. Proses ini bertujuan untuk menyederhanakan dataset dan memastikan bahwa hanya fitur yang bermakna yang digunakan dalam tahap analisis selanjutnya.

4.1.1 *Data Acquisition*

Penelitian ini menggunakan data dari repositori GitHub StatsBomb open-data yang diambil melalui proses unduhan menggunakan *tool* aria2 di Google Colab dengan bahasa pemrograman Python. Untuk mengakses data tersebut, pertama-tama dilakukan pengunduhan *file master.zip* yang berisi data pertandingan sepak bola. Berikut adalah tahapan yang dilakukan dalam proses pengumpulan data:

- a. Mengunduh *file* master.zip dari repositori GitHub StatsBomb open-data dengan menggunakan perintah di Google Colab dan *tool* aria2 untuk mempercepat proses unduhan.
- b. Menyusun skrip Python di Google Colab untuk mengekstrak semua *event* data yang ada pada file yang diunduh.
- c. Mengkonversi *event* data menjadi format *dataframe* menggunakan *pandas* untuk mempermudah pengolahan data lebih lanjut.
- d. Menyimpan *dataframe* yang telah diproses dalam format *parquet* untuk memudahkan analisis data selanjutnya. *Dataframe* yang dihasilkan mencakup berbagai kolom terkait informasi pertandingan, yang akan diseleksi dan diproses lebih lanjut untuk analisis yang lebih mendalam.

4.1.2 Pemilihan Kompetisi

Langkah ini bertujuan untuk memastikan bahwa hanya data pertandingan yang relevan dan sesuai dengan fokus penelitian yang digunakan dalam proses analisis. Data mentah yang tersedia di repositori open-data StatsBomb mencakup berbagai jenis kompetisi, termasuk pertandingan pria, wanita, dan kelompok usia muda. Oleh karena itu, proses seleksi dilakukan secara sistematis untuk menyaring data berdasarkan dua kriteria utama: jenis kelamin peserta dan tingkat kompetisi.

Data kompetisi difilter untuk hanya menyertakan pertandingan pria dengan memeriksa atribut *competition_gender* yang bernilai '*male*'. Selanjutnya, untuk memastikan bahwa hanya kompetisi tingkat senior yang disertakan, dilakukan pengecualian terhadap kompetisi yang mengandung kata kunci seperti 'U21', 'U23', 'U18', dan lainnya dalam nama kompetisi, yang menunjukkan kelompok usia muda.

Setelah mendapatkan daftar kompetisi yang valid, data pertandingan (*matches*) dari kompetisi tersebut dimuat dan difilter lebih lanjut untuk hanya menyertakan pertandingan antara dua tim pria. Proses ini menghasilkan kumpulan data pertandingan yang sesuai dengan fokus penelitian, yaitu analisis pertandingan sepak bola pria tingkat senior. Tabel 4.1 Menunjukkan Daftar Kompetisi yang akan digunakan.

Tabel 4.1 Daftar Kompetisi

<i>Competition ID</i>	<i>Season ID</i>	<i>Competition Name</i>
11	4	FIFA World Cup
2	44	Premier League
37	90	La Liga
72	30	UEFA Champions League
43	106	Bundesliga
49	3	Serie A
4	1	Ligue 1
55	27	Copa America
9	42	African Cup of Nations
16	1	Eredivisie

4.2 Data Preprocessing

Tahap *preprocessing* adalah tahapan yang berisi serangkaian proses untuk membersihkan dan menyiapkan data agar siap digunakan dalam analisis dan pemodelan pada tahapan selanjutnya. Dengan *preprocessing* yang tepat, kualitas data meningkat dan hasil pemodelan di tahap berikutnya menjadi lebih akurat dan andal.

4.2.1 Pemilihan Data

Pemilihan jenis *event* yang tepat sangat krusial untuk memastikan relevansi dan kualitas analisis. Berdasarkan dokumentasi resmi dari StatsBomb, setiap peristiwa dalam pertandingan dikategorikan dengan identifier unik. *Event* dengan *type.id* 16 merepresentasikan aksi "*Shot*" atau tembakan, yang menjadi fokus utama dalam model xG karena langsung berkaitan dengan upaya mencetak gol.

Setelah data kompetisi kita dapatkan, maka selanjutnya kita akan mengambil data *event* yang sesuai dengan kompetisi yang kita dapatkan dengan mendapatkan *match_id* yang akan digunakan nantinya. Kemudian kita akan menyimpan semua *event* yang sesuai dengan *match_id* pada *file parquet*. Selanjutnya kita dapat memilih *event shot* yang ada pada setiap pertandingan untuk dijadikan *dataset* untuk pelatihan dan pengujian model nantinya. Hasil *dataframe* ditunjukkan pada Gambar 4.1.

period	minute	second	start_x	start_y	team_name	player_name	end_x	end_y	type	...
1	7	15	115.4	29.4	England	Harry Maguire	120.0	34.9	16	...
1	26	58	101.1	55.3	England	Bukayo Saka	117.5	41.9	16	...
1	29	8	113.4	49.1	England	Mason Mount	120.0	45.0	16	...
1	31	47	110.5	40.7	England	Harry Maguire	120.0	36.8	16	...
1	34	9	112.0	38.0	England	Jude Bellingham	120.0	43.0	16	...

Gambar 4.1 Hasil Pemilihan Data

4.2.2 Pemilihan Kolom

Pada tahapan ini, proses awal yang dilakukan adalah memilih sub set data yang sesuai dengan tujuan penelitian. Dalam konteks prediksi xG, hanya data

dengan tipe *event Shot* yang diambil karena data tersebut merepresentasikan momen-momen tembakan yang menjadi fokus analisis. Setelah *event Shot* teridentifikasi, dilakukan proses seleksi kolom atau fitur yang dianggap memiliki nilai prediktif terhadap hasil tembakan (*shot outcome*). Fitur-fitur yang dipilih mencakup atribut spasial (seperti posisi awal dan akhir tembakan), temporal (menit dan detik), teknis (teknik tembakan, bagian tubuh yang digunakan), serta konteks permainan (tekanan lawan, pola permainan, dan tipe *event* sebelumnya). Fitur-fitur ini ditujukan untuk menangkap berbagai aspek yang dapat memengaruhi kemungkinan sebuah tembakan menjadi gol. Hasil seleksi kolom ditampilkan pada Tabel 4.2.

Tabel 4.2 Nama dan Deskripsi Kolom

Nama Kolom	Deskripsi
<i>period</i>	Periode pertandingan saat tembakan terjadi
<i>minute</i>	Menit pertandingan saat tembakan dilakukan
<i>second</i>	Detik pertandingan saat tembakan dilakukan
<i>start_x, start_y</i>	Koordinat awal tembakan (lokasi pemain saat menembak)
<i>position</i>	Posisi pemain di dalam tim (Bek, Gelandang, Penyerang)
<i>shot_outcome</i>	Hasil tembakan (0 = tidak gol, 1 = gol)
<i>shot_body_part</i>	Bagian tubuh yang digunakan dalam menembak
<i>shot_first_time</i>	Apakah tembakan dilakukan secara langsung tanpa kontrol bola
<i>shot_one_on_one</i>	Apakah tembakan dilakukan dalam situasi satu lawan satu dengan kiper
<i>shot_open_goal</i>	Apakah tembakan dilakukan ke gawang yang kosong
<i>shot_aerial_won</i>	Apakah pemain memenangkan duel udara sebelum tembakan
<i>shot_key_pass</i>	Apakah tembakan didahului oleh umpan kunci
<i>possession</i>	Nomor penguasaan bola dari tim

<i>play_pattern</i>	Pola permainan yang terjadi sebelum tembakan
<i>under_pressure</i>	Apakah pemain berada dalam tekanan saat melakukan tembakan
<i>shot_technique</i>	Teknik tembakan yang digunakan

Pemilihan variabel ini bertujuan untuk menyederhanakan kompleksitas data serta meningkatkan fokus pada fitur-fitur yang relevan dalam konteks perhitungan xG. Langkah ini dilakukan untuk mengurangi redundansi informasi dan meminimalkan risiko *overfitting* akibat penggunaan variabel yang tidak informatif. Selain itu, untuk fitur-fitur yang bersifat kategorial, dilakukan pendekatan dengan mengambil langsung nilai ID atau representasi numerik yang sudah tersedia dari masing-masing kategori. Dengan demikian, proses ini menghindari kebutuhan akan transformasi tambahan seperti *one-hot encoding* atau *label encoding*, yang dapat menambah dimensi data secara signifikan tanpa memberikan kontribusi informatif yang sepadan. Pendekatan ini tidak hanya menjaga efisiensi pemrosesan data, tetapi juga mempertahankan struktur semantik dari variabel kategorial dalam bentuk yang lebih ringkas dan langsung digunakan oleh model. Gambar 4.2 menunjukkan contoh data setelah proses pemilihan variabel dilakukan.

	period	minute	second	location_x	...	technique
0	1	1	42	111.0	...	93
1	1	4	47	96.0	...	93
2	1	8	37	107.0	...	93
3	1	17	26	111.0	...	93
4	1	21	16	105.0	...	93
...
68863	2	65	39	106.1	...	93
68864	2	69	0	114.9	...	93
68865	2	82	41	103.6	...	93
68866	2	85	10	108.5	...	93
68867	2	85	58	117.0	...	91

Gambar 4.2 Contoh Data Sesudah Pemilihan Variabel

4.2.3 Data Cleansing

Tahap data *cleansing* dilakukan untuk memeriksa kelengkapan dan keunikan data dengan tujuan memastikan bahwa tidak terdapat nilai kosong (*missing values*) maupun data duplikat yang dapat memengaruhi proses analisis. Pada penelitian ini, proses pembersihan data menunjukkan bahwa data yang digunakan telah bersih secara struktural. Hal ini disebabkan oleh karakteristik data sepak bola yang cenderung unik di mana setiap peristiwa dalam pertandingan memiliki identitas dan konteks yang berbeda serta karena data yang disediakan oleh StatsBomb telah tersusun secara rapi dan konsisten. Struktur data yang baik ini sangat membantu dalam mempercepat proses *preprocessing* dan meningkatkan kualitas hasil analisis, karena tidak memerlukan upaya koreksi data secara signifikan.

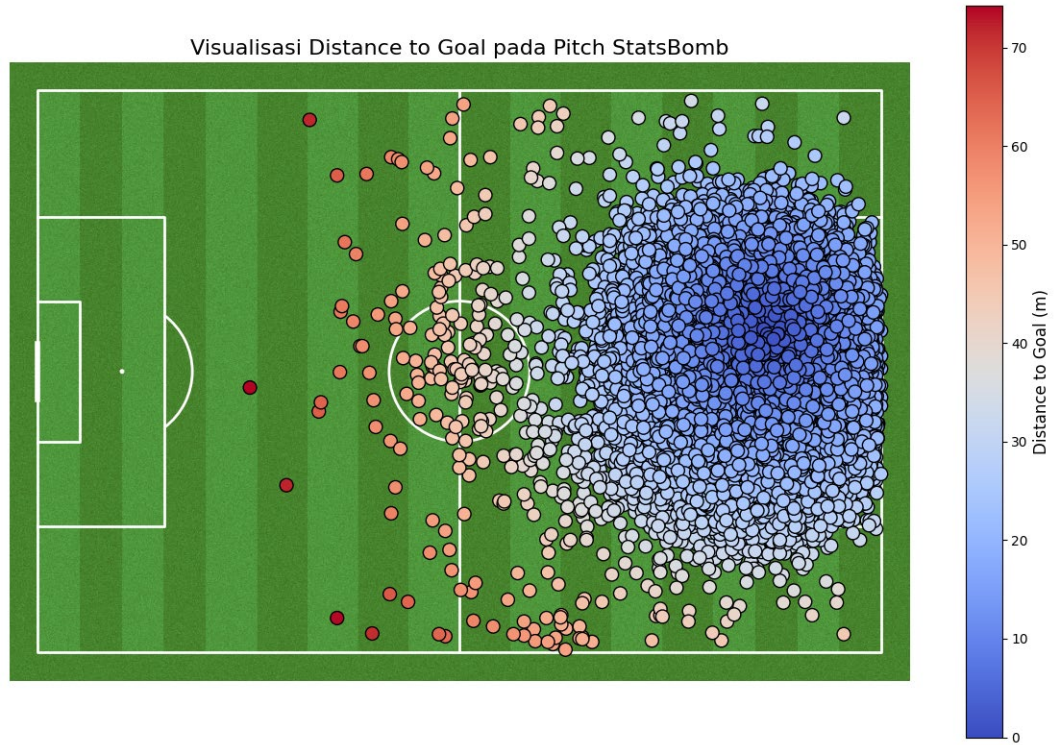
4.3 Data Transformation

4.3.1 Feature Engineering

Tahapan pertama dalam proses *transformation* pada penelitian ini adalah melakukan *feature engineering* dengan menambahkan tiga fitur baru, yaitu jarak dan sudut tembakan terhadap gawang serta kejadian sebelum terjadinya *shot*. Fitur ini ditambahkan untuk memberikan informasi spasial yang lebih kaya kepada model, mengingat lokasi dan sudut tembakan serta momentum sangat berpengaruh terhadap kemungkinan terciptanya gol.

a. *Distance to Goal*

Fitur pertama dalam proses *feature engineering* adalah menghitung jarak antara posisi tembakan dan pusat gawang. Informasi spasial ini penting karena jarak tembakan merupakan salah satu faktor utama yang memengaruhi kemungkinan terciptanya gol. Semakin dekat jarak tembakan ke gawang, secara umum peluang untuk mencetak gol menjadi lebih besar. Gambar 4.4 menunjukkan visualisasi fitur *distance to goal* pada lapangan pertandingan berdasarkan koordinat StatsBomb. Titik-titik pada visualisasi merepresentasikan lokasi awal tembakan, dengan warna yang menunjukkan jaraknya terhadap gawang, semakin biru berarti semakin dekat dan semakin merah berarti semakin jauh.



Gambar 4.3 Visualisasi *Distance to Goal*

Dalam *dataset* ini, koordinat pusat gawang StatsBomb berada pada titik ($x = 104.0$, $y = 34.0$), yang merepresentasikan titik tengah di antara dua tiang gawang. Jarak dihitung menggunakan rumus *Euclidean distance*, yaitu akar kuadrat dari jumlah kuadrat selisih antara koordinat tembakan dan koordinat pusat gawang. Secara matematis, perhitungan ini dinyatakan sebagai berikut:

$$Distance = \sqrt{(x_{goal} - x_{start})^2 + (y_{goal} - y_{start})^2} \quad (4.1)$$

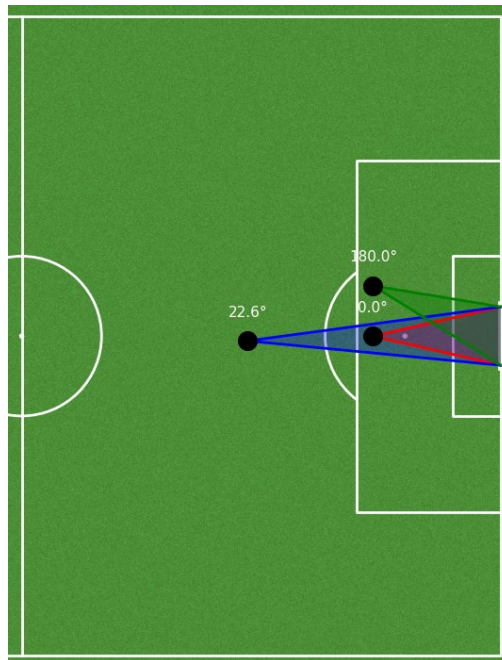
Fungsi ini diimplementasikan dalam kode Python yang akan mengembalikan nilai jarak dalam satuan relatif terhadap sistem koordinat StatsBomb. Hasil perhitungan disimpan dalam kolom baru bernama *distance_to_goal* dan digunakan sebagai salah satu fitur masukan dalam model prediksi. Gambar 4.7 menunjukkan contoh hasil dari proses penambahan fitur ini.

distance_to_goal
15.448625
14.905368
26.828716
10.837435
19.568598

Gambar 4.4 *Distance to Goal*

b. *Angle to Goal*

Selain jarak, fitur penting lainnya yang ditambahkan dalam proses *feature engineering* adalah sudut tembakan terhadap gawang, atau dikenal sebagai *open play angle*. Fitur ini merepresentasikan seberapa besar ruang terbuka yang tersedia bagi penembak untuk mengarahkan bola ke area di antara kedua tiang gawang. Semakin lebar sudut yang terbuka, semakin besar peluang tembakan untuk menghasilkan gol.



Gambar 4.5 Visualisasi Sudut Tembakan

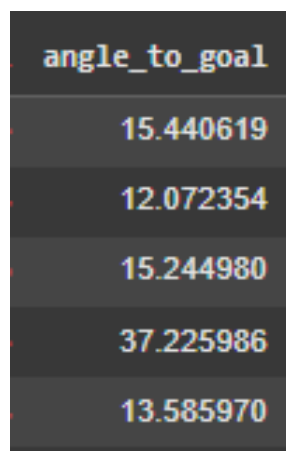
Perhitungan sudut dilakukan dengan mengacu pada tiga titik, posisi tembakan dan dua titik tiang gawang (kanan dan kiri). Dalam sistem koordinat StatsBomb, gawang terletak pada posisi horizontal tetap yaitu $x = 120$, dengan tiang bawah (kanan) berada pada $y = 43,66$ dan tiang atas (kiri) pada $y = 36,34$. Fungsi python digunakan untuk menghitung besar sudut terbuka menggunakan hukum cosinus. Langkah-langkahnya meliputi:

- i) Hitung jarak dari posisi tembakan ke masing-masing tiang gawang (A dan B).
- ii) Hitung panjang sisi antara kedua tiang (C).
- iii) Gunakan hukum cosinus untuk mencari sudut di antara kedua sisi tersebut.

Sudut dalam radian dikonversi ke derajat menggunakan fungsi `np.degrees()`.

Jika tembakan dilakukan tepat dari titik tengah gawang ($x = 120$ dan y berada

antara dua tiang), maka sudut maksimal akan diberikan sebesar 180 derajat. Sebaliknya, jika posisi tembakan berada sejajar secara horizontal dengan gawang tetapi tidak dalam rentang vertikal antara tiang, maka sudut dianggap 0 derajat. Hasil perhitungan ini disimpan dalam kolom *angle_to_goal*, yang menjadi input penting dalam proses pemodelan. Gambar 4.5 menunjukkan hasil dari *feature engineering* tersebut.



angle_to_goal
15.440619
12.072354
15.244980
37.225986
13.585970

Gambar 4.6 *Angle to Goal*

c. *Type Before*

Fitur *type_before* ditambahkan sebagai bagian dari proses *feature engineering* untuk memberikan konteks temporal terhadap peristiwa tembakan yang dianalisis. Fitur ini merepresentasikan jenis *event* yang terjadi tepat sebelum tembakan dilakukan, dengan mengambil nilai *type.id* dari *event* sebelumnya dalam urutan kronologis pertandingan. Informasi ini bertujuan untuk menangkap dinamika permainan yang mendahului tembakan, seperti apakah tembakan tersebut terjadi setelah dribel, operan, intersepsi, atau aksi defensif lawan. Tabel 4.3 beberapa *type.id* umum dalam data, yang digunakan untuk mengidentifikasi jenis peristiwa dalam pertandingan sepak bola.

Tabel 4.3 Deskripsi Jenis *type* dalam Pertandingan Sepak Bola.

Event Type	Type ID	Deskripsi Singkat
<i>50/50</i>	33	Dua pemain dari tim berbeda berebut bola lepas.
<i>Bad Behaviour</i>	24	Pelanggaran di luar permainan yang berujung kartu.
<i>Ball Receipt*</i>	42	Momen penerimaan atau usaha menerima operan.
<i>Ball Recovery</i>	2	Usaha merebut kembali bola lepas.
<i>Block</i>	6	Pemain menghalangi bola dengan tubuhnya.
<i>Carry</i>	43	Pemain menguasai bola saat bergerak atau diam.
<i>Clearance</i>	9	Menghalau bola dari area bahaya tanpa niat mengoper ke rekan.
<i>Dispossessed</i>	3	Pemain kehilangan bola karena ditekel tanpa mencoba dribel.
<i>Dribble</i>	14	Usaha pemain melewati lawan dengan menggiring bola.
<i>Dribbled Past</i>	39	Pemain dilewati oleh lawan saat dribel.
<i>Duel</i>	4	Duel 1v1 antara pemain dari tim berbeda.
<i>Error</i>	37	Kesalahan pemain yang mengarah pada tembakan lawan.
<i>Foul Committed</i>	22	Pelanggaran yang dilakukan terhadap lawan (tidak termasuk <i>offside</i>).
<i>Foul Won</i>	21	Pelanggaran yang diterima dan menghasilkan tendangan bebas atau penalti.
<i>Goal Keeper</i>	23	Segala aksi penjaga gawang (<i>penyelamatan, smother, punch, dll</i>).
<i>Half End</i>	34	Peluit akhir babak pertandingan oleh wasit.
<i>Half Start</i>	18	Peluit awal babak pertandingan oleh wasit.
<i>Injury Stoppage</i>	40	Penghentian permainan karena cedera.
<i>Interception</i>	10	Pemain memotong jalur operan lawan untuk mencegah bola sampai ke target.
<i>Miscontrol</i>	38	Kehilangan kontrol bola karena sentuhan yang buruk.
<i>Offside</i>	8	Pelanggaran posisi <i>offside</i> .

<i>Own Goal Against</i>	20	Gol bunuh diri oleh tim sendiri.
<i>Own Goal For</i>	25	Gol bunuh diri yang menguntungkan tim.
<i>Pass</i>	30	Umpan dari satu pemain ke pemain lain.
<i>Player Off</i>	27	Pemain keluar lapangan tanpa pergantian (misalnya karena cedera).
<i>Player On</i>	26	Pemain kembali masuk ke lapangan setelah <i>Player Off</i> .
<i>Pressure</i>	17	Aksi menekan pemain lawan di area tertentu, direkam bersama durasi tekanan.
<i>Referee Ball-Drop</i>	41	Wasit menjatuhkan bola untuk melanjutkan pertandingan setelah jeda (misalnya cedera).
<i>Shield</i>	28	Pemain melindungi bola agar keluar lapangan tanpa dikejar lawan.
<i>Shot</i>	16	Upaya mencetak gol dengan bagian tubuh legal.
<i>Starting XI</i>	35	Informasi awal pemain yang bermain dan formasi tim.
<i>Substitution</i>	19	Pergantian pemain saat pertandingan berlangsung.
<i>Tactical Shift</i>	36	Perubahan posisi pemain atau formasi taktik dalam pertandingan.

Dengan menambahkan konteks ini, model dapat memahami alur permainan yang berujung pada tembakan dan mengenali pola peristiwa yang secara statistik lebih mungkin menghasilkan gol. Fitur *type_before* diisi hanya jika terdapat *event* sebelumnya, jika tembakan merupakan *event* pertama dalam urutan, maka fitur ini dikosongkan. Gambar 4.7 menunjukkan hasil dari *feature engineering* tersebut.

type_before
2
43
42
4
4

Gambar 4.7 Type Before

4.3.2 Seleksi Fitur

Tahap ini bertujuan untuk memilih fitur-fitur yang paling relevan dan berpengaruh terhadap prediksi model, sehingga dapat meningkatkan efisiensi dan akurasi pemodelan. Seleksi fitur dilakukan setelah proses *feature engineering* selesai, dengan mempertimbangkan konteks domain serta performa masing-masing fitur dalam mendukung prediksi *shot_outcome*. Fitur yang memiliki kontribusi kecil atau *redundan* dapat dihilangkan untuk menghindari kompleksitas berlebih dan mengurangi risiko *overfitting*. Proses ini membantu model fokus pada informasi yang benar-benar penting. Tabel 4.1 menunjukkan fitur-fitur yang dipertahankan setelah melalui tahap seleksi.

Tabel 4.4 Fitur-Fitur Pada Tahap Seleksi

No.	Fitur
1	<i>minute</i>
2	<i>second</i>
3	<i>play_pattern</i>
4	<i>position</i>
5	<i>shot_technique</i>

6	<i>shot_body_part</i>
7	<i>shot_type</i>
8	<i>shot_first_time</i>
9	<i>shot_open_goal</i>
10	<i>shot_one_on_one</i>
11	<i>shot_aerial_won</i>
12	<i>under_pressure</i>
13	<i>distance_to_goal</i>
14	<i>angle_to_goal</i>
15	<i>shot_key_pass</i>
16	<i>start_x</i>
17	<i>start_y</i>
18	<i>possession</i>

4.3.3 Pemisahan Data Uji dan Data Latih

Salah satu tahapan penting dalam proses *transformation* adalah pemisahan data menjadi data latih dan data uji. Tujuan dari proses ini adalah untuk mengevaluasi kinerja model secara objektif terhadap data yang belum pernah digunakan dalam proses pelatihan. Pemisahan data dilakukan menggunakan fungsi *train_test_split* dari *library scikit-learn*, dengan proporsi 90% data sebagai data latih dan 10% sebagai data uji. Parameter *random_state* disetel ke angka 42 untuk menjamin konsistensi hasil pemisahan saat kode dijalankan ulang. Setelah proses ini dilakukan, diperoleh 61.981 baris data untuk pelatihan dan 6.887 baris data untuk pengujian. Gambar 4.6 menunjukkan jumlah baris dan kolom data latih dan uji.

	Train	Test
Rows	61981	6887
Columns	17	17

Gambar 4.8 Jumlah Baris dan Kolom Data Latih dan Uji.

4.4 Data Mining

4.4.1 Perancangan Model

Pada penelitian ini, algoritma yang digunakan untuk membangun model adalah LightGBM, yang dikenal memiliki efisiensi tinggi dan performa unggul pada data dengan dimensi besar.

a. Inisiasi Model

Langkah pertama dalam perancangan model adalah melakukan inisialisasi algoritma yang akan digunakan. Proses inisialisasi dilakukan dengan membuat objek *LGBMClassifier* dari *library scikit-learn*, dan parameter *random_state* diatur ke nilai 42 untuk memastikan hasil yang *reproduksibel*. Selanjutnya, ditentukan ruang pencarian *hyperparameter* yang akan digunakan dalam proses *tuning*, yang meliputi parameter *min_child_samples*, *num_leaves*, *reg_lambda*, *reg_alpha*, dan *max_depth*. Parameter-parameter ini diatur dalam bentuk distribusi acak menggunakan fungsi *sp_randint* dan *sp_uniform* dari *scipy.stats*, yang akan menjadi acuan dalam proses pemilihan kombinasi terbaik pada tahap pencarian *hyperparameter* berikutnya.

b. Definisi Fungsi Scoring

Setelah inisiasi model dilakukan, langkah selanjutnya adalah mendefinisikan fungsi penilaian yang digunakan untuk mengevaluasi performa model selama

proses *hyperparameter tuning*. Pada penelitian ini digunakan dua metrik evaluasi, yaitu ROC AUC dan *Brier Score*. Fungsi penilaian ini didefinisikan dalam bentuk *dictionary* dengan nama *scoring*, di mana ROC AUC dipanggil secara langsung dan *Brier Score* didefinisikan menggunakan *make_scorer* dari *scikit-learn* dengan argumen *greater_is_better=False* karena nilai *Brier Score* yang lebih kecil menunjukkan performa yang lebih baik.

c. Pelatihan Model

Setelah fungsi penilaian ditentukan, tahap berikutnya adalah menginisialisasi proses *hyperparameter tuning* menggunakan *RandomizedSearchCV*. Teknik ini digunakan untuk mencari kombinasi parameter terbaik dari model LightGBM berdasarkan evaluasi dengan *5-fold cross-validation* dan 100 iterasi pencarian. Tidak seperti proses *tuning* standar yang hanya mempertimbangkan satu metrik, dalam penelitian ini digunakan pendekatan *refit kustom* untuk menyeimbangkan antara akurasi klasifikasi dan kualitas kalibrasi probabilitas.

Fungsi *custom_refit* yang digunakan bertujuan untuk memilih model dengan nilai ROC AUC tertinggi, tetapi juga mempertimbangkan *Brier Score* agar model yang terpilih tidak hanya mampu mengklasifikasi dengan baik, tetapi juga memberikan estimasi probabilitas yang kalibrasinya baik. Jika model dengan ROC AUC tertinggi memiliki *Brier Score* lebih kecil dari ambang batas tertentu (misalnya -0.1), maka model tersebut akan dipilih. Jika tidak, sistem akan memilih model dengan *Brier Score* terbaik. Setelah objek *RandomizedSearchCV* dikonfigurasi dengan parameter, metrik penilaian, dan

fungsi *refit*, proses pelatihan model dilakukan dengan memanggil fungsi *fit* pada data latih (*X_train* dan *y_train*).

Parameter-parameter ini dipilih secara otomatis oleh algoritma pencarian (*RandomizedSearchCV*) berdasarkan kinerja terbaik dalam pelatihan. Setiap parameter memainkan peran penting dalam mengontrol struktur pohon, regularisasi, pembobotan, serta teknik pembelajaran untuk meningkatkan akurasi dan mencegah *overfitting*. Tabel 4.5 menunjukkan konfigurasi akhir model LightGBM setelah *tuning*.

Tabel 4.5 Konfigurasi Akhir Model LightGBM Setelah *Tuning*

Parameter	Nilai
cv	3
method	isotonic
boosting_type	gbdt
num_leaves	15
max_depth	84
min_child_samples	146
min_child_weight	0.001
min_split_gain	0.0
colsample_bytree	1.0
subsample	1.0
subsample_for_bin	200000
subsample_freq	0
learning_rate	0.1
n_estimators	100
reg_alpha	0.513

reg_lambda	0.971
random_state	42
importance_type	split

d. Pemilihan Model Terbaik dan Kalibrasi Probabilitas

Setelah proses *RandomizedSearchCV* selesai, langkah selanjutnya adalah mengambil model terbaik yang diperoleh dari hasil pencarian *hyperparameter*. Model terbaik ini diakses melalui atribut *best_estimator_* dan merupakan konfigurasi LightGBM dengan performa optimal berdasarkan kriteria *custom refit* yang telah ditentukan sebelumnya. Untuk meningkatkan akurasi estimasi probabilitas, dilakukan tahap kalibrasi menggunakan *CalibratedClassifierCV* dengan metode *isotonic regression* dan *cross-validation* sebanyak tiga lipatan. Kalibrasi ini bertujuan agar *output* probabilitas dari model lebih merefleksikan tingkat kepercayaan yang sebenarnya, terutama dalam konteks prediksi tembakan yang menghasilkan gol atau tidak. Proses pelatihan ulang dilakukan terhadap model yang telah dikalibrasi menggunakan data latih sebelum model digunakan dalam tahap evaluasi akhir.

4.4.2 Permodelan LGBM

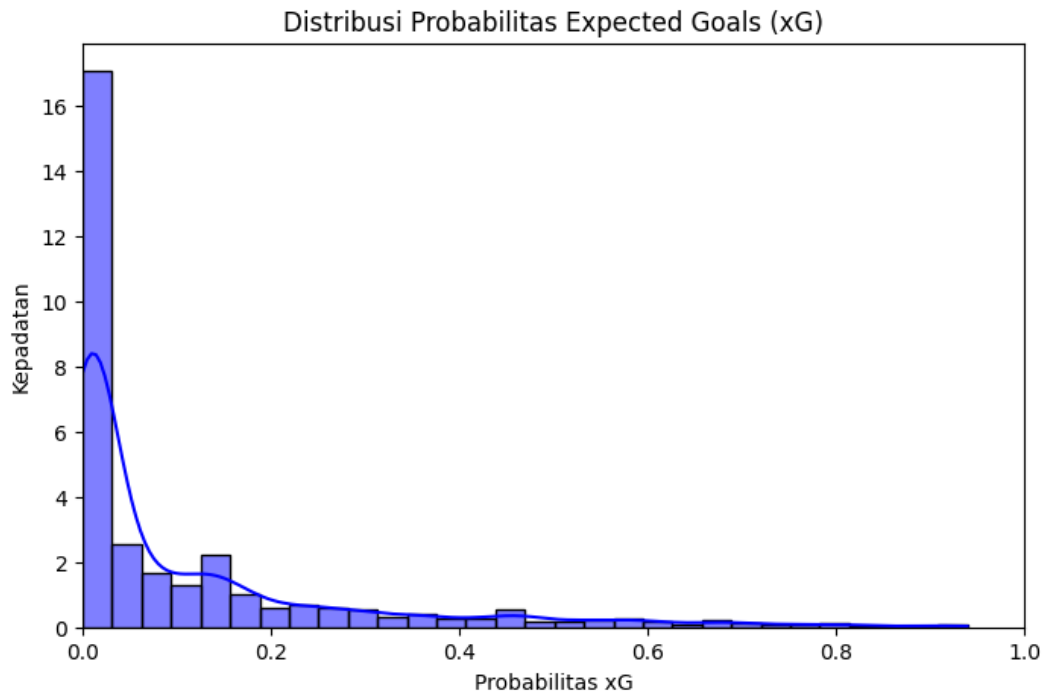
Setelah melakukan pencarian nilai *hyperparameter* terbaik melalui proses *RandomizedSearchCV*, nilai-nilai tersebut digunakan pada tahap permodelan akhir. Nilai-nilai parameter pada *hyperparameter* mencerminkan peran penting dari dua teknik inti dalam LightGBM, yaitu *Gradient-based One-Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB).

Teknik GOSS memungkinkan model untuk fokus pada *instance* dengan gradien tinggi, yang biasanya lebih informatif untuk proses pembelajaran, sehingga mempercepat pelatihan tanpa mengorbankan akurasi. Hal ini sangat relevan dengan parameter *min_child_samples* yang cukup besar (146), karena membantu menjaga kestabilan pembagian *node* meskipun jumlah data yang di sampling dikurangi oleh GOSS. Sementara itu, teknik EFB menggabungkan fitur-fitur eksklusif yang tidak aktif bersamaan, sehingga memungkinkan penggunaan jumlah *num_leaves* yang besar (15) tanpa meningkatkan kompleksitas model secara drastis.

4.4.3 Visualisasi

Salah satu langkah awal dalam proses evaluasi model prediktif adalah dengan memahami pola distribusi dari nilai xG yang dihasilkan. Dalam konteks ini, dilakukan visualisasi berupa histogram dan kurva distribusi (KDE plot) terhadap nilai-nilai prediksi dari model LGBM. Tujuan utama dari visualisasi ini adalah untuk mengidentifikasi karakteristik sebaran data, termasuk kecenderungan *skewness*, serta untuk memahami apakah model cenderung menghasilkan prediksi yang konservatif (nilai xG rendah) atau agresif (nilai xG tinggi). Gambar 4.9 menyajikan histogram distribusi nilai xG yang diprediksi oleh model LGBM, dilengkapi dengan estimasi *kernel density estimation* (KDE) untuk memberikan

representasi yang lebih halus terhadap bentuk sebaran tersebut.



Gambar 4.9 Histogram Distribusi Nilai xG

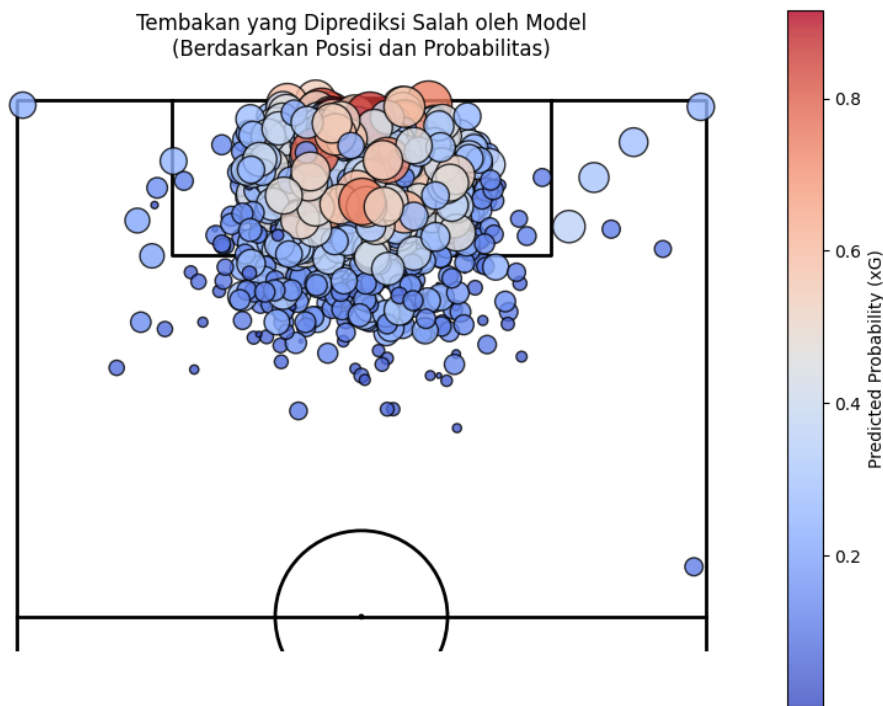
Distribusi nilai xG yang dihasilkan oleh model LGBM menunjukkan karakteristik *positively skewed* yang sangat kuat, dengan sebagian besar nilai terkonsentrasi mendekati 0.0, sebagaimana ditunjukkan oleh frekuensi tertinggi pada bin pertama (sekitar 0.0–0.02) dan puncak tajam kurva KDE di titik tersebut. Hal ini mencerminkan realitas bahwa mayoritas tembakan dalam sepak bola merupakan peluang berkualitas rendah misalnya, tembakan dari jarak jauh atau sudut sempit dengan probabilitas gol yang sangat kecil. Modus distribusi yang berada sangat dekat dengan nol memperkuat interpretasi ini. Di sisi lain, distribusi juga memiliki *long tail* ke kanan, yang membentang hingga nilai xG tinggi (seperti 0.4, 0.6, hingga mendekati 1.0), merepresentasikan peluang emas seperti penalti, tap-in, atau situasi satu lawan satu yang meskipun jarang terjadi, menyumbang

proporsi signifikan terhadap total peluang gol. Pola ini mencerminkan distribusi *heavy-tailed* atau bahkan menyerupai *power law*, di mana sebagian kecil peristiwa ekstrem (tembakan berkualitas sangat tinggi) memberikan dampak besar terhadap akumulasi nilai xG total. Kurva KDE berperan penting dalam menghaluskan bentuk distribusi dan memperjelas pola yang tidak tampak secara eksplisit dalam histogram, termasuk punuk kecil di kisaran xG menengah (0.15–0.4) yang dapat menunjukkan keberadaan sub-kategori peluang dengan tingkat kesulitan sedang.

Area di bawah kurva KDE antara dua titik (misalnya 0.05–0.10) dapat diinterpretasikan sebagai probabilitas kumulatif kemunculan tembakan dalam rentang tersebut, dengan luas area yang sangat besar di dekat 0.0 menunjukkan bahwa probabilitas tembakan memiliki xG rendah ($P(0 \leq xG \leq 0.05)$) sangat tinggi. Secara keseluruhan, distribusi ini mencerminkan sifat permainan sepak bola yang berkarakter skor rendah, dengan dominasi tembakan risiko rendah dan hanya sebagian kecil peluang berkualitas tinggi, sehingga memperkuat validitas statistik serta kesesuaian konteks model dengan realitas permainan.

Melanjutkan dari analisis distribusi numerik, pendekatan visual berbasis spasial juga digunakan untuk memperdalam interpretasi terhadap perilaku model dalam konteks pertandingan sebenarnya. Peta tembakan (shot map) disusun untuk memvisualisasikan lokasi-lokasi di lapangan tempat tembakan dilakukan, dengan setiap titik pada peta dilengkapi oleh nilai xG yang diprediksi oleh model. Representasi visual ini menggunakan variasi warna atau ukuran titik sebagai indikator kuantitatif nilai xG, sehingga memungkinkan identifikasi cepat terhadap tembakan-tembakan berisiko tinggi maupun rendah berdasarkan posisi

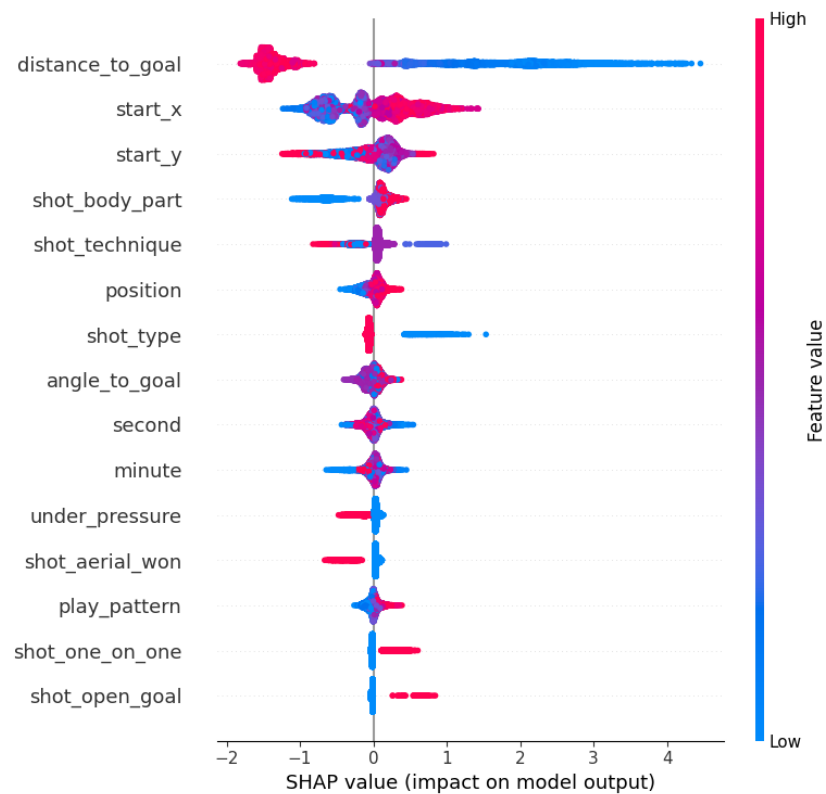
geografisnya di dalam area permainan. Gambar 4.10 menyajikan peta sebaran tembakan tersebut, yang menjadi alat bantu penting dalam memahami dinamika spasial peluang dalam suatu pertandingan secara intuitif dan informatif.



Gambar 4.10 Peta Sebaran dengan Nilai xG

Selain visualisasi spasial melalui peta tembakan, interpretasi lebih mendalam terhadap perilaku internal model dilakukan dengan menggunakan pendekatan explainable AI, yaitu SHAP (*SHapley Additive exPlanations*). Metode ini memungkinkan analisis kontribusi masing-masing fitur terhadap prediksi nilai xG, baik secara global (menjelaskan pengaruh fitur terhadap keseluruhan prediksi model) maupun secara lokal (menjelaskan kontribusi fitur terhadap prediksi spesifik pada satu observasi/tembakan). Dengan memanfaatkan SHAP, dapat diidentifikasi fitur mana yang paling berperan dalam membentuk nilai prediksi, seperti jarak tembakan, sudut terhadap gawang, jenis aksi sebelum tembakan, atau

posisi pemain bertahan terdekat. Visualisasi pada Gambar 4.11 menunjukkan bagaimana nilai SHAP terdistribusi untuk berbagai fitur penting dalam model, serta bagaimana masing-masing nilai input memengaruhi naik atau turunnya prediksi xG.



Gambar 4.11 Visualisasi SHAP untuk Interpretasi Model xG

Analisis SHAP global menunjukkan *distance_to_goal* sebagai fitur paling berpengaruh, dengan nilai rendah (titik biru) memberikan SHAP positif tinggi yang mendorong prediksi gol, dan nilai tinggi (titik merah) memberikan SHAP negatif yang menurunkan peluang gol, mencerminkan hubungan invers yang sangat kuat dan intuitif. Fitur kedua terpenting, *start_x*, memperlihatkan nilai tinggi (posisi lebih maju di lapangan) terkait SHAP positif, menunjukkan tendangan dari posisi lebih dekat gawang lawan meningkatkan peluang gol, sedangkan nilai rendah

terkait SHAP negatif. Sedangkan *start_y* memiliki pengaruh signifikan, nilai ekstrem di sisi lapangan (titik merah dan biru) memberikan SHAP negatif atau mendekati nol, sementara nilai tengah lapangan menghasilkan SHAP positif, mengindikasikan bahwa tendangan dari posisi sentral lebih berpeluang gol karena sudut tembak yang lebih menguntungkan dibandingkan posisi samping yang sempit.

Fitur *shot_body_part* menunjukkan dampak cukup besar, dengan nilai tinggi (mewakili kaki dominan atau sundulan) berkorelasi positif terhadap prediksi gol, sedangkan nilai rendah (kaki non-dominan atau bagian tubuh lain) cenderung negatif, menandakan efektivitas bagian tubuh tertentu dalam mencetak gol. *shot_technique* juga berpengaruh signifikan, teknik tendangan tertentu (seperti *volleys* atau tendangan melengkung) memiliki SHAP positif, sedangkan teknik lain mengurangi peluang gol. Pada *position*, nilai SHAP positif terkait posisi pemain menyerang (*striker*, *midfielder*), sementara posisi bertahan (bek, kiper) cenderung negatif, mengindikasikan posisi sangat memengaruhi probabilitas gol. *shot_type* memperlihatkan tipe tendangan efektif (tendangan kuat, penalti) dengan nilai SHAP positif, berbeda dengan tendangan spekulatif yang negatif. Sedangkan *angle_to_goal* memiliki pengaruh moderat, sudut tembak lebar (nilai tinggi) menghasilkan SHAP positif, sedangkan sudut sempit negatif, sesuai dengan logika bahwa sudut tembak lebih terbuka meningkatkan peluang gol.

Fitur waktu seperti *second* dan *minute* berpengaruh kecil, dengan distribusi SHAP yang tersebar dan pola halus, menunjukkan waktu dalam detik atau menit saat tendangan diambil hanya memiliki efek minimal atau spesifik terhadap

probabilitas gol. *under_pressure* menunjukkan bahwa tendangan dalam tekanan (titik merah) memiliki SHAP negatif, sedangkan tanpa tekanan positif, konsisten dengan penurunan peluang gol saat ada pengawalan ketat. *shot_aerial_won* memiliki dampak kecil, duel udara yang dimenangkan sebelum tendangan (titik merah) memberikan sedikit peningkatan probabilitas gol. *play_pattern* juga berdampak kecil, dengan pola tertentu (serangan balik cepat, *set piece*) memberikan SHAP positif, dan pola lain (*open play* stagnan) negatif, mengindikasikan peran pola serangan terhadap efektivitas tembakan.

Fitur situasional seperti *shot_one_on_one* memiliki pengaruh kecil namun konsisten, tendangan dalam situasi satu lawan satu dengan kiper (titik merah) meningkatkan probabilitas gol, sedangkan bukan situasi tersebut cenderung netral atau negatif. Terakhir, *shot_open_goal* memiliki dampak paling kecil tetapi sangat positif, menunjukkan bahwa tendangan ke gawang kosong secara jelas meningkatkan probabilitas gol walaupun frekuensinya rendah, dan model secara efektif menangkap efek signifikan ini.

4.5 Evaluation

Tahap evaluasi dilakukan terhadap data uji yang telah dipisahkan pada proses data *transformation*. Evaluasi bertujuan untuk mengukur sejauh mana model LightGBM mampu memberikan prediksi probabilistik yang akurat terhadap kemungkinan terciptanya gol (*expected goals*). Metrik evaluasi yang digunakan pada penelitian ini terdiri dari dua metrik utama, yaitu *Brier Score* dan ROC AUC. Kedua metrik ini dipilih karena sesuai dengan tujuan dari model xG, yaitu

menghasilkan prediksi dalam bentuk probabilitas, bukan klasifikasi biner semata. Tabel 4.6 berikut menunjukkan hasil evaluasi model LightGBM berdasarkan kedua metrik tersebut.

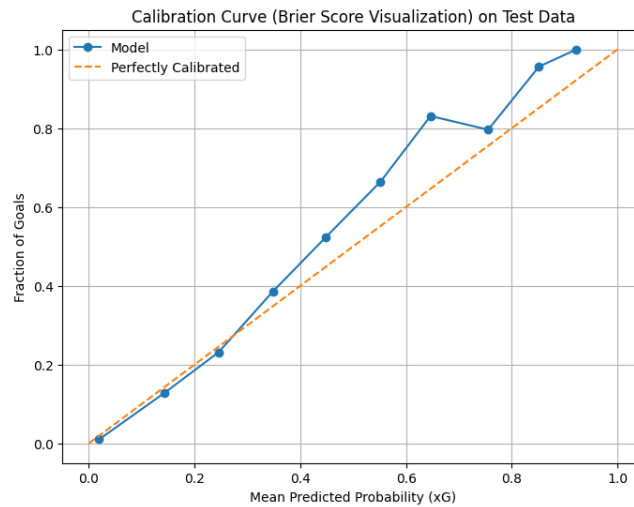
Tabel 4.6 Hasil Evaluasi Model LightGBM

Metrik Evaluasi	Nilai
<i>Brier Score</i>	0.0663
ROC AUC	0.9134

4.5.2 *Brier Score*

Nilai *Brier Score* sebesar 0.0663 mengindikasikan bahwa prediksi probabilitas yang dihasilkan oleh model memiliki tingkat kesalahan kuadrat yang sangat rendah. Hal ini menandakan bahwa model mampu mengestimasi peluang terciptanya gol secara akurat dan konsisten terhadap data aktual pada data uji. Dalam konteks model *expected goals* (xG), nilai *Brier Score* yang rendah sangat penting karena model ini tidak hanya bertujuan untuk mengklasifikasi hasil tembakan, tetapi juga memberikan probabilitas yang merepresentasikan peluang realistis terjadinya gol.

Visualisasi hasil kalibrasi pada data uji ditampilkan pada Gambar 4.12, yang menunjukkan hubungan antara rata-rata probabilitas prediksi (xG) dan proporsi aktual terjadinya gol (*fraction of goals*). Kurva yang mendekati garis diagonal membuktikan bahwa model memiliki kalibrasi yang baik dan dapat diandalkan dalam memberikan prediksi probabilitas.



Gambar 4.12 Calibration Curve (Brier Score Visualization) on Test Data

Lebih lanjut, nilai ini menunjukkan bahwa prediksi probabilitas yang dikeluarkan oleh model memiliki kalibrasi yang baik, artinya semakin tinggi nilai probabilitas prediksi suatu tembakan, maka semakin besar pula kecenderungannya untuk benar-benar menjadi gol dalam data aktual. Ini dibuktikan melalui kurva kalibrasi, di mana garis biru (hasil model) mengikuti dengan cukup dekat garis oranye putus-putus yang merepresentasikan kondisi kalibrasi sempurna. Meskipun terdapat sedikit deviasi pada beberapa titik terutama di rentang probabilitas menengah ke atas namun secara keseluruhan, distribusi prediksi tetap mencerminkan tren empiris yang logis dan stabil.

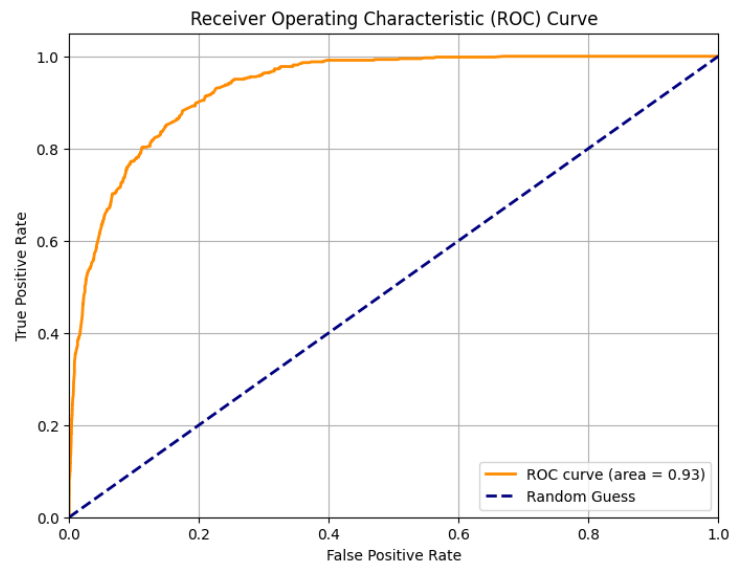
Kinerja ini menjadi indikator bahwa proses kalibrasi menggunakan metode *isotonic regression* berhasil menyesuaikan *output* model sehingga prediksi probabilitas tidak bersifat *overconfident* maupun *underconfident*. Hal ini sangat krusial pada skenario data *imbalanced* seperti prediksi xG, di mana sebagian besar

tembakan memang tidak berujung pada gol, dan model cenderung rawan bias terhadap mayoritas kelas.

4.5.3 ROC AUC

Model LightGBM menunjukkan performa diskriminatif yang sangat baik dengan nilai ROC AUC sebesar 0.93. Nilai ini menunjukkan bahwa model memiliki kemampuan yang sangat tinggi dalam membedakan antara tembakan yang menghasilkan gol dan tembakan yang tidak. ROC *curve* yang dihasilkan memiliki kurva yang menjauh signifikan dari garis diagonal (*random guess*), mengindikasikan bahwa model mampu mengidentifikasi *true positive* dengan tingkat *false positive* yang rendah pada sebagian besar *threshold*.

Performa ini juga memperlihatkan bahwa fitur-fitur yang digunakan dalam pelatihan, termasuk informasi spasial, teknik tembakan, serta konteks permainan (seperti tekanan dan posisi lawan), berhasil dikombinasikan secara efektif oleh model untuk mengenali pola-pola yang berkontribusi terhadap terciptanya gol. ROC *curve* yang stabil dan area kurva yang luas merupakan indikator bahwa model dapat digunakan dalam skenario nyata yang memerlukan prediksi *ranking* (misalnya untuk mengurutkan kualitas peluang tembakan). Visualisasi kinerja diskriminatif model ditampilkan pada Gambar 4.13, yang memperlihatkan perbandingan antara ROC *curve* model dengan *baseline random guess*.



Gambar 4.13 Receiver Operating Characteristic (ROC) Curve

4.5.4 Perbandingan dengan Model Pada Literatur Lain

Untuk menilai kinerja model yang dikembangkan dalam penelitian ini secara lebih komprehensif, dilakukan perbandingan terhadap hasil evaluasi dari beberapa model xG yang telah dikembangkan pada studi sebelumnya. Hasil dari model LightGBM yang dibangun dalam penelitian ini menunjukkan kinerja yang sangat kompetitif, dengan *Brier Score* sebesar 0.0663 dan ROC AUC sebesar 0.9134, mengungguli sebagian besar model yang ada dalam literatur terdahulu. Rangkuman perbandingan hasil evaluasi model pada beberapa studi terdahulu dapat dilihat pada Tabel 4.7.

Tabel 4.7 Perbandingan Hasil Evaluasi Model xG pada Berbagai Literatur

Penulis & Tahun	Model	<i>Brier Score</i>	ROC AUC
Scholtes & Karakuş (2024)	<i>Bayesian Hierarchical</i>	0.075	–
ElHabr (2023)	XGBoost (Opta npxG)	0.0715	–
Cavus & Biecek (2022)	LightGBM	0.173	0.818
Haaren (2021)	<i>Boosting Machine</i>	0.082	0.793

Eggels et al. (2016)	<i>Random Forest</i>	–	0.814
Anzer & Bauer (2021)	GBM	–	0.822
Mead et al. (2023)	XGBoost	0.0799	0.800
Model Penelitian ini	LightGBM + Calibration	0.0663	0.913

Berdasarkan Tabel 4.7, model LightGBM yang dikembangkan dalam penelitian ini menunjukkan kinerja superior dibandingkan model-model xG yang telah dikembangkan sebelumnya. Dengan Brier Score sebesar 0.0663 dan ROC AUC 0.9134, model ini mencatatkan hasil terbaik di antara model pembanding.

Model ini mengungguli pendekatan Bayesian hierarchical dari Scholtes & Karakuş (2024) dengan Brier Score 0.075, meskipun ROC AUC tidak dilaporkan. Demikian pula, model ini lebih unggul dibandingkan XGBoost dari Opta npxG perusahaan sports analytics asal Inggris menurut ElHabr (2023) yang mencatat Brier Score 0.0715, meskipun tanpa nilai ROC AUC. Sementara itu, model LightGBM Cavus & Biecek (2022) menghasilkan Brier Score 0.173 dan ROC AUC 0.818, jauh di bawah model ini meskipun menggunakan algoritma yang sama.

Model ini juga lebih baik dibandingkan XGBoost Mead et al. (2023) dengan Brier Score 0.0799 dan ROC AUC 0.800, serta model Boosting Machine dari Haaren (2021) yang mencatat Brier 0.082 dan ROC AUC 0.793. Model klasik seperti Random Forest (Eggels et al., 2016) dan GBM (Anzer & Bauer, 2021) masing-masing mencatat ROC AUC 0.814 dan 0.822, namun tanpa pelaporan Brier Score.

4.6 Interpretasi Hasil

Model expected goals (xG) yang dikembangkan dalam penelitian ini menunjukkan performa yang cukup baik dalam mengestimasi probabilitas terjadinya gol berdasarkan variabel-variabel yang relevan dalam situasi tembakan. Evaluasi performa model melalui metrik-metrik seperti Brier score, log loss, dan ROC-AUC mengindikasikan bahwa model memiliki tingkat akurasi prediksi yang memadai serta kestabilan yang layak untuk digunakan dalam konteks analisis performa sepak bola. Hal ini menunjukkan bahwa model mampu menangkap pola-pola signifikan dalam data historis dan menerjemahkannya menjadi prediksi probabilistik yang representatif terhadap kemungkinan terciptanya gol.

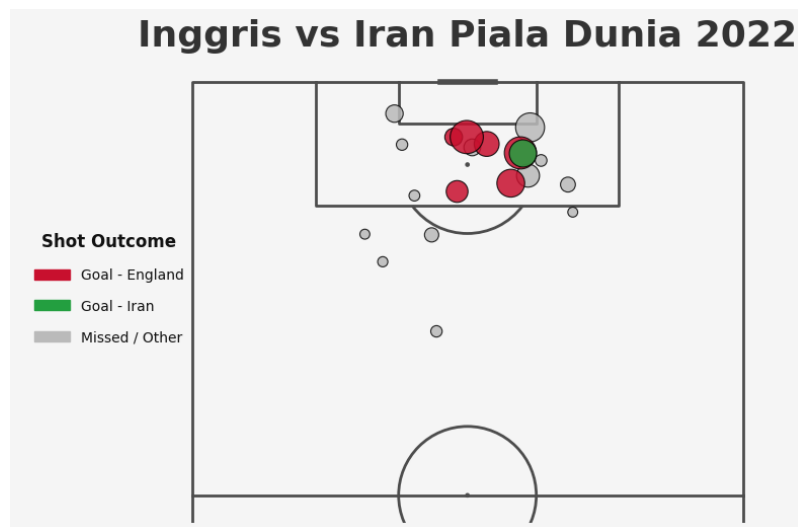
Untuk menginterpretasikan lebih lanjut kemampuan model dalam konteks dunia nyata, dilakukan penerapan perhitungan xG terhadap data spesifik dari suatu pertandingan sebagai studi kasus. Sebagai contoh, Tabel 4.8 menyajikan visualisasi distribusi nilai xG dari masing-masing peluang yang tercipta dalam pertandingan antara tim nasional Inggris melawan Iran pada Piala Dunia FIFA 2022. Hasil ini memberikan gambaran empiris mengenai kualitas peluang yang diciptakan oleh masing-masing tim selama pertandingan berlangsung, serta menunjukkan bagaimana model dapat digunakan untuk menganalisis efisiensi konversi peluang menjadi gol secara lebih objektif dan terukur.

Tabel 4.8 Statistik Efektivitas Penyerangan Berdasarkan *Expected Goals*

Tim Nasional	Total <i>Expected Goals</i> (xG)	Jumlah Tembakan	Jumlah Gol Aktual	Rata-rata xG per Tembakan	Diferensial Gol (Gol – xG)
Inggris	3.09	13	6	0.238	+2.91
Iran	0.87	7	1	0.125	+0.13

Selain dari sisi akurasi prediktif, model xG yang dikembangkan juga memberikan kontribusi signifikan dalam konteks visualisasi dan analisis pertandingan sepak bola. Dengan kemampuannya mengkuantifikasi kualitas peluang secara numerik, model ini dapat menghasilkan metrik xG yang informatif dan aplikatif, sehingga dapat menjadi alat bantu strategis bagi tim analis dalam mengevaluasi performa tim maupun pemain secara lebih objektif. Analisis berbasis xG juga membuka ruang untuk interpretasi yang lebih dalam terhadap dinamika pertandingan, tidak hanya berdasarkan skor akhir, tetapi juga berdasarkan kualitas peluang yang diciptakan dan dihadapi.

Lebih lanjut, model ini juga dapat diintegrasikan sebagai alat ukur performa yang tervisualisasi secara intuitif dan informatif. Gambar 4.14 menyajikan visualisasi sebaran tembakan (shot map) beserta nilai xG masing-masing peluang yang terjadi pada pertandingan antara Inggris melawan Iran di Piala Dunia 2022. Visualisasi ini memungkinkan pemahaman spasial yang lebih baik terhadap lokasi dan kualitas peluang, serta membantu mengidentifikasi area-area strategis yang menjadi sumber utama ancaman serangan selama pertandingan. Dengan demikian, model xG ini tidak hanya berfungsi sebagai alat prediksi, tetapi juga sebagai media analisis taktis yang kaya informasi.



Gambar 4.14 *Shot Map* Inggris vs Iran

Untuk mengevaluasi sejauh mana performa model xG ini dalam konteks prediksi pertandingan secara langsung, dilakukan perbandingan hasil prediksi dengan data xG yang disediakan oleh beberapa penyedia statistik sepak bola ternama seperti Opta, Pro Football Focus (PFF), FBref, dan xGScore. Perbandingan ini dilakukan pada tiga pertandingan kunci di ajang Piala Dunia 2022, yaitu Inggris vs Iran, Inggris vs Prancis, dan Argentina vs Kroasia, serta satu pertandingan final UEFA Euro 2024. Tabel 4.9 menyajikan nilai xG yang dihasilkan oleh model ini pada masing-masing pertandingan tersebut, beserta perbandingannya dengan estimasi dari penyedia statistik lainnya. Analisis ini bertujuan untuk menilai konsistensi dan validitas model dalam konteks aplikatif, serta menakar sejauh mana model yang dikembangkan mampu menghasilkan estimasi yang kompetitif dibandingkan dengan standar industri dalam bidang analisis sepak bola berbasis data.

Tabel 4.9 Perbandingan Model dengan Penyedia Statistik Sepak Bola

<i>Pertandingan</i>	<i>Skor</i>	<i>Sumber</i>	<i>xG Tim A</i>	<i>xG Tim B</i>
---------------------	-------------	---------------	-----------------	-----------------

<i>England vs Iran – WC 2022</i>	6 – 2	LGBM	England: 3.09	Iran: 0.87
		Opta	England: 2.109	Iran: 1.751
		xGScore.io	England: 2.14	Iran: 1.42
		FBref	England: 2.1	Iran: 1.4
		PFF	England: 2.14	Iran: 1.62
<i>England vs France – WC 2022</i>	1 – 2	LGBM	England: 1.98	France: 0.64
		PFF	England: 2.4	France: 0.73
		xGScore.io	England: 2.55	France: 1.21
		FBref	England: 2.4	France: 0.9
		Opta	England: 2.407	France: 1.012
<i>Argentina vs Croatia – WC 2022</i>	3 – 0	LGBM	Argentina: 2.01	Croatia: 0.95
		PFF	Argentina: 2.12	Croatia: 0.30
		xGScore.io	Argentina: 2.76	Croatia: 0.57
		Opta	Argentina: 2.33	Croatia: 0.52
		FBref	Argentina: 2.3	Croatia: 0.5
<i>Spain vs England – Final EURO 2024</i>	2 – 1	LGBM	England: 0.67	Spain: 1.63
		xGScore.io	England: 0.63	Spain: 1.9
		FBref	England: 0.5	Spain: 1.9
		Opta	England: 0.527	Spain: 1.953
		PFF	–	–

Pada pertandingan antara Inggris melawan Iran di fase grup Piala Dunia 2022 yang berakhir dengan skor 6–2, model yang dikembangkan dalam penelitian ini menghasilkan estimasi nilai xG sebesar 3,09 untuk Inggris dan 0,87 untuk Iran. Jika dibandingkan dengan data dari penyedia statistik lainnya, terdapat perbedaan yang cukup signifikan. Opta mencatat xG sebesar 2,109 (Inggris) dan 1,751 (Iran), sementara xgscore.io melaporkan nilai 2,14 (Inggris) dan 1,42 (Iran). FBref

memberikan estimasi serupa yaitu 2,1 untuk Inggris dan 1,4 untuk Iran, sedangkan PFF mencatat 2,14 untuk Inggris dan 1,62 untuk Iran. Meskipun terdapat variasi antar penyedia, model ini menunjukkan kecenderungan yang lebih tinggi dalam memperkirakan dominasi Inggris, dengan nilai xG yang mencerminkan secara lebih jelas disparitas kualitas peluang yang tercipta di antara kedua tim.

Pada pertandingan perempat final antara Inggris dan Prancis (1–2), model ini memprediksi xG sebesar 1,98 untuk Inggris dan 0,64 untuk Prancis. Angka ini mengindikasikan bahwa Inggris menciptakan peluang dengan kualitas lebih tinggi dibanding Prancis, meskipun hasil akhir menunjukkan sebaliknya. Bila dibandingkan dengan penyedia data lainnya, PFF mencatat 2,4 (Inggris) dan 0,73 (Prancis), xgscore.io memberikan 2,55 dan 1,21, sementara FBref dan Opta masing-masing memperkirakan 2,4 dan 0,9 serta 2,407 dan 1,012. Secara umum, model ini memberikan estimasi yang lebih konservatif untuk Prancis, namun tetap sejalan dengan kesimpulan bahwa Inggris memiliki dominasi peluang dalam pertandingan tersebut.

Selanjutnya, pada laga semifinal antara Argentina dan Kroasia yang berakhir dengan kemenangan Argentina 3–0, model ini memperkirakan nilai xG sebesar 2,01 untuk Argentina dan 0,95 untuk Kroasia. Estimasi ini relatif sejalan dengan hasil observasi dan mendekati beberapa penyedia data resmi. PFF mencatat 2,12 untuk Argentina dan hanya 0,30 untuk Kroasia. Sementara itu, xgscore.io dan FBref memberikan nilai yang sedikit lebih tinggi untuk Argentina, yakni masing-masing 2,76 dan 2,3, dan nilai lebih rendah untuk Kroasia, yaitu 0,57 dan 0,5. Opta memberikan estimasi sebesar 2,336 (Argentina) dan 0,520 (Kroasia). Secara umum,

model ini menampilkan prediksi yang stabil dan mencerminkan keseimbangan realistis antara dominasi Argentina dan ketidakmampuan Kroasia menciptakan peluang berkualitas.

Terakhir, pada pertandingan final Euro 2024 antara Spanyol dan Inggris yang berakhir dengan kemenangan Spanyol 2–1, model ini menghasilkan nilai xG sebesar 1,63 untuk Spanyol dan 0,67 untuk Inggris. Estimasi ini mendekati angka dari beberapa penyedia statistik. Xgscore.io mencatat nilai sebesar 1,90 (Spanyol) dan 0,63 (Inggris), sementara FBref dan Opta memberikan hasil serupa, yakni 1,9 dan 0,5 (FBref), serta 1,953 dan 0,527 (Opta). Sayangnya, data dari PFF untuk pertandingan ini tidak tersedia. Perbandingan ini menunjukkan bahwa model yang dikembangkan mampu memberikan prediksi yang sejalan dengan tren umum yang tercermin dalam data statistik publik, sehingga memperkuat validitas model sebagai alat analisis performa pertandingan sepak bola tingkat tinggi.

4.7 Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, keterbatasan utama terletak pada kelengkapan dan cakupan data. Dataset yang digunakan berasal dari sumber open-data StatsBomb yang hanya mencakup liga dan turnamen tertentu, seperti Liga Inggris, La Liga, dan Piala Dunia. Hal ini membatasi kemampuan generalisasi model terhadap kompetisi lain yang memiliki karakteristik permainan berbeda, baik dari segi level kompetisi, gaya bermain, maupun kualitas pemain. Selain itu, meskipun data StatsBomb dikenal kaya akan detail teknis, sejumlah fitur krusial dalam analisis xG seperti posisi kiper atau

intensitas tekanan dari pemain bertahan tidak selalu tersedia atau hanya tersedia dalam jumlah terbatas (misalnya *freeze frame* data). Kekosongan ini dapat mengurangi kemampuan model dalam merepresentasikan konteks situasional dari suatu tembakan secara menyeluruh.

Kedua, keberlakuan model yang dibangun secara spesifik pada data dari satu liga atau turnamen tertentu dapat membatasi performanya ketika diterapkan pada kompetisi lain. Perbedaan gaya bermain antar liga, taktik dominan, tingkat kemampuan teknis pemain, serta kondisi permainan yang kontekstual dapat memengaruhi performa model secara signifikan. Dengan demikian, validitas eksternal dari model ini masih perlu diuji secara lebih luas sebelum dapat digunakan secara general.

Ketiga, keterbatasan dalam pemahaman domain atau domain knowledge turut menjadi tantangan dalam eksplorasi fitur. Tanpa pemahaman mendalam mengenai peran spesifik pemain, strategi taktis, dan pola permainan, terdapat kemungkinan bahwa beberapa fitur bersifat terlalu dangkal atau bahkan mengarah pada interpretasi yang menyesatkan. Hal ini menunjukkan pentingnya kolaborasi antara peneliti data dan praktisi atau analis sepak bola untuk memperkaya proses *feature engineering* dan interpretasi model.