BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian terdahulu merupakan bagian penting dalam persiapan awal melakukan penelitian mendalam. Penelitian terdahulu memberikan gambaran dan perkembangan ilmiah dalam lingkup prediksi pemain sepak bola. Penting untuk meninjau penelitian-penelitian sebelumnya guna mengidentifikasi metode yang digunakan, jenis datanya, hasil utama yang diperoleh, serta kesenjangan penelitian yang masih dapat dikembangkan lebih lanjut. Untuk tinjauan terhadap karya-karya penelitian terdahulu ini yang berupa jurnal akademik, studi dipilih berdasarkan kesesuaiannya dengan isu-isu dalam lingkup penilian pemain sepak bola, penerapan teknik machine learning, relevansinya dengan model regresi, serta keterkaitannya dengan aspek pengambilan keputusan terkait proses rekrutmen pemain.

Secara umum, penelitian mengenai valuasi pemain sepak bola dalam beberapa tahun terakhir menunjukkan pergeseran dari pendekatan penilaian tradisional menuju pendekatan prediktif berbasis data. Nilai pasar pemain tidak lagi hanya dipahami sebagai hasil penilaian subjektif, tetapi sebagai keluaran ekonomi yang dapat dianalisis melalui atribut pemain, performa teknis, potensi, popularitas, serta konteks pasar. Oleh karena itu, pendekatan machine learning semakin banyak digunakan karena mampu menangkap hubungan kompleks dan nonlinier antara berbagai faktor yang memengaruhi nilai pemain. Ringkasan penelitian terdahulu yang relevan dengan penelitian ini disajikan pada Tabel 2.1.

Tabel 2. 1 Tabel Penelitian Terdahulu

Penelitian Terdahulu 1	
Komponen	**Isi**
Judul	A Novel Machine Learning Method for Estimating Football Players' Value in the Transfer Market
Penulis	Behravan & Razavi
Sumber jurnal	Soft Computing, Vol. 25, No. 3, pp. 2499–2511, Springer.
Tahun	2021
Permasalahan	Penelitian ini membahas kebutuhan estimasi nilai pemain sepak bola secara lebih objektif menggunakan pendekatan

	machine learning, karena nilai pemain dalam pasar transfer dipengaruhi oleh banyak atribut dan sulit diprediksi hanya dengan pendekatan konvensional.
Framework / Metode	Menggunakan APSO clustering untuk mengelompokkan pemain secara otomatis dan PSO-SVR untuk membangun model regresi. PSO digunakan untuk seleksi fitur dan optimasi parameter SVR.
Data yang digunakan	Dataset FIFA 20.
Temuan	Metode clustering otomatis membagi pemain menjadi empat klaster, yaitu penjaga gawang, gelandang, pemain bertahan, dan penyerang. Model menghasilkan estimasi nilai pemain dengan akurasi sebesar 74%, serta PSO menunjukkan performa lebih baik dibandingkan GWO, IPO, dan WOA.
Pembahasan	Penelitian ini menunjukkan bahwa data FIFA dapat digunakan untuk memperkirakan nilai pemain sepak bola. Selain itu, penggunaan metode machine learning mampu membantu proses estimasi nilai pemain berdasarkan atribut yang tersedia dalam dataset.
Gap Penelitian	Penelitian tersebut belum membandingkan secara langsung Random Forest Regression dan XGBoost Regression. Selain itu, dataset yang digunakan masih FIFA 20, sedangkan penelitian ini menggunakan FIFA 24 Player Stats Dataset. Penelitian tersebut juga belum menekankan evaluasi pada dua skala, yaitu skala logaritmik dan skala nilai pasar asli.
Relevansi terhadap penelitian ini	Relevan karena sama-sama membahas prediksi nilai pasar pemain sepak bola menggunakan data FIFA dan pendekatan machine learning. Penelitian tersebut menjadi dasar bahwa atribut pemain dalam dataset FIFA dapat digunakan untuk membangun model estimasi nilai pemain.
Penelitian Terdahulu 2	
Komponen	**Isi**
Judul	*Predicting Transfer Fees in Professional European Football Before and During COVID-19 Using Machine Learning*
Penulis	Yang, Koenigstorfer, & Pawlowski
Sumber jurnal	*European Sport Management Quarterly*, Vol. 24, No. 3, pp. 603–623, Routledge.
Tahun	2021
Permasalahan	Penelitian ini membahas prediksi biaya transfer pemain sepak bola profesional Eropa sebelum dan selama pandemi COVID-19, karena kondisi pasar transfer mengalami perubahan dan dapat memengaruhi nilai transaksi pemain.

Framework / Metode	Menggunakan Random Forest estimator, Generalized Additive Models, dan Quantile Additive Models untuk menganalisis hubungan non-linear antara variabel pemain dan biaya transfer.
Data yang digunakan	Data Transfermarkt dengan jumlah 3.512 transfer pemain pada periode sebelum dan selama pandemi COVID-19. Target yang digunakan adalah biaya transfer pemain.
Temuan	Beberapa prediktor biaya transfer memiliki hubungan non-linear. Model yang dilatih menggunakan data sebelum COVID-19 cenderung memprediksi biaya transfer selama pandemi terlalu rendah, terutama pada pemain dengan nilai transfer menengah dan tinggi.
Pembahasan	Penelitian ini menunjukkan bahwa pendekatan machine learning non-linear dapat memberikan pemahaman lebih baik terhadap dinamika pasar transfer pemain, khususnya ketika terjadi perubahan kondisi pasar.
Gap Penelitian	Penelitian tersebut berfokus pada transfer fee, bukan market value. Selain itu, penelitian tersebut tidak secara khusus membandingkan Random Forest Regression dan XGBoost Regression untuk memprediksi nilai pasar pemain berdasarkan FIFA 24 Player Stats Dataset.
Relevansi terhadap penelitian ini	Relevan karena menunjukkan bahwa model non-linear seperti Random Forest dapat digunakan untuk memahami pola nilai dalam pasar transfer sepak bola. Namun, penelitian ini berbeda karena berfokus pada prediksi nilai pasar pemain, bukan biaya transfer aktual.
Penelitian Terdahulu 3	
Komponen	**Isi**
Judul	Developing A Predictive Model for Football Players Market Value Using Machine Learning
Penulis	Idris & Ng
Sumber jurnal	Journal of Informatics and Web Engineering, Vol. 4, No. 3, pp. 203–214.
Tahun	2025
Permasalahan	Penelitian ini membahas kebutuhan model prediktif untuk memperkirakan nilai pasar pemain sepak bola berdasarkan data performa pemain dari liga-liga Eropa
Framework / Metode	Menggunakan Random Forest, LightGBM, XGBoost, dan GDBT dengan tahapan preprocessing, feature selection, encoding, standarisasi, serta hyperparameter tuning menggunakan RandomizedSearchCV.
Data yang digunakan	Data statistik performa pemain dari lima liga top Eropa pada tahun 2017–2020.
Temuan	Penelitian membangun model prediksi nilai pasar pemain dan mengidentifikasi fitur performa yang memiliki korelasi terhadap nilai pasar. Namun, detail model optimal dan

	metrik evaluasi tidak dijelaskan secara eksplisit dalam sumber yang digunakan
Pembahasan	Penelitian ini menunjukkan bahwa beberapa algoritma ensemble dapat digunakan dalam prediksi nilai pasar pemain. Tahapan preprocessing dan pemilihan fitur juga menjadi bagian penting sebelum proses pemodelan.
Gap Penelitian	Penelitian tersebut belum memberikan penjelasan metrik model secara lengkap dalam sumber yang digunakan. Selain itu, penelitian ini belum secara khusus menekankan perbandingan Random Forest Regression dan XGBoost Regression pada FIFA 24 Player Stats Dataset dengan evaluasi log dan skala asli.
Relevansi terhadap penelitian ini	Relevan karena sama-sama menggunakan pendekatan machine learning untuk memprediksi nilai pasar pemain sepak bola dan menggunakan beberapa algoritma ensemble sebagai pembanding
Penelitian Terdahulu 4	
Komponen	**Isi**
Judul	Predicting the Value of Football Players: Machine Learning Techniques and Sensitivity Analysis Based on FIFA and Real-World Statistical Datasets
Penulis	Shen
Sumber jurnal	Applied Intelligence, Vol. 55, Article No. 265, Springer.
Tahun	2025
Permasalahan	Penelitian ini membahas prediksi nilai pemain sepak bola dengan menggabungkan data FIFA dan data statistik nyata agar estimasi nilai pemain lebih akurat dan dapat dijelaskan.
Framework / Metode	Menggunakan SVR, Random Forest Regression, XGBoost, CatBoost, model ansambel berbasis Dempster-Shafer Theory, SHAP, dan FAST.
Data yang digunakan	Dataset FIFA 19 dan real-world statistical dataset yang mencakup metrik performa, atribut fisiologis, dan variabel kontekstual pemain
Temuan	XGBoost yang dioptimasi menggunakan SCSO menghasilkan kesalahan estimasi sekitar 1,7 juta dolar, sedangkan RFR, SVR, CatBoost, dan XGBoost lain berada sekitar 2 juta dolar. Model ansambel dinilai mampu menghasilkan estimasi nilai pemain yang lebih akurat.
Pembahasan	Penelitian ini menunjukkan bahwa model ensemble dan teknik interpretabilitas dapat meningkatkan kualitas prediksi nilai pemain. Penggunaan data FIFA juga terbukti relevan untuk penelitian valuasi pemain.
Gap Penelitian	Penelitian tersebut menggunakan pendekatan yang lebih kompleks melalui model ansambel dan analisis sensitivitas, sedangkan penelitian ini lebih fokus pada perbandingan

	langsung Random Forest Regression dan XGBoost Regression. Selain itu, penelitian ini menggunakan FIFA 24 Player Stats Dataset dan menekankan interpretasi hasil sebagai pendukung keputusan rekrutmen.
Relevansi terhadap penelitian ini	Relevan karena membahas prediksi nilai pemain menggunakan FIFA dataset, Random Forest, dan XGBoost. Penelitian ini dapat menjadi pembanding bahwa algoritma ensemble cocok digunakan untuk kasus valuasi pemain.
Penelitian Terdahulu 5	
Komponen	**Isi**
Judul	*Forecasting the Future Development in Quality and Value of Professional Football Players*
Penulis	van Arem, Goes-Smit, & Söhl
Sumber jurnal	*Applied Sciences*, Vol. 15, No. 16, Article No. 8916, MDPI.
Tahun	2025
Permasalahan	Penelitian ini membahas prediksi perkembangan kualitas dan nilai pemain sepak bola profesional di masa depan
Framework / Metode	Menggunakan OLS, Lasso, LME, Decision Tree, Random Forest, XGBoost, dan kNN.
Data yang digunakan	Data SciSports, yaitu data kualitas pemain tahun 2014–2022 dan data nilai pemain tahun 2016–2021
Temuan	Random Forest dinilai paling sesuai untuk memprediksi nilai pemain dengan kesalahan di atas 2,8 juta.
Pembahasan	Penelitian ini menunjukkan bahwa Random Forest mampu memberikan performa baik dalam prediksi nilai pemain dibandingkan beberapa metode lain.
Gap Penelitian	Penelitian tersebut berfokus pada forecasting perkembangan kualitas dan nilai pemain, sedangkan penelitian ini berfokus pada prediksi nilai pasar pemain berdasarkan atribut FIFA 24. Selain itu, penelitian ini membandingkan Random Forest dan XGBoost secara khusus pada evaluasi log dan USD.
Relevansi terhadap penelitian ini	Relevan karena menunjukkan bahwa Random Forest dapat digunakan untuk memprediksi nilai pemain sepak bola dan dapat dibandingkan dengan model lain seperti XGBoost.
Penelitian Terdahulu 6	
Komponen	**Isi**
Judul	*When Interpretable Machine Learning Meets the Beautiful Game: A Predictive Analytics Approach to Soccer Player Valuation in the Transfer Market*
Penulis	Li et al
Sumber jurnal	Sport, Business and Management: An International Journal, Vol. 16, No. 2, pp. 240–261, Emerald Publishing Limited.

Tahun	2025
Permasalahan	Penelitian ini membahas prediksi valuasi pemain sepak bola dalam pasar transfer dengan pendekatan machine learning yang dapat diinterpretasikan.
Framework / Metode	Menggunakan Decision Tree, Random Forest, SVR, dan XGBoost, serta interpretabilitas model menggunakan SHAP.
Data yang digunakan	831 data transfer pemain pada tahun 2017–2020 yang bersumber dari Transfermarkt.
Temuan	XGBoost menjadi model terbaik dengan nilai RMSE sebesar 0,717 dan R^2 sebesar 0,695. Atribut penting yang ditemukan antara lain sisa kontrak, usia, dan peringkat tim.
Pembahasan	Penelitian ini menunjukkan bahwa XGBoost memiliki kemampuan baik dalam prediksi valuasi pemain dan interpretabilitas model dapat membantu menjelaskan faktor yang mempengaruhi nilai pemain.
Gap Penelitian	Penelitian tersebut menggunakan data Transfermarkt dan menekankan interpretabilitas melalui SHAP, sedangkan penelitian ini menggunakan FIFA 24 Player Stats Dataset. Selain itu, penelitian ini tidak hanya melihat performa model pada satu skala, tetapi juga mengevaluasi hasil pada skala logaritmik dan skala asli nilai pasar.
Relevansi terhadap penelitian ini	Relevan karena menggunakan Random Forest dan XGBoost dalam konteks valuasi pemain sepak bola. Penelitian ini juga mendukung pentingnya interpretasi hasil model dalam pengambilan keputusan
Penelitian Terdahulu 7	
Komponen	**Isi**
Judul	*Dynamic Financial Valuation of Football Players: A Machine Learning Approach Across Career Stages*
Penulis	Khalife et al.
Sumber jurnal	International Journal of Financial Studies, Vol. 13, No. 2, Article No. 111, MDPI.
Tahun	2025
Permasalahan	Penelitian ini membahas valuasi finansial pemain sepak bola secara dinamis dengan mempertimbangkan perbedaan tahap karier, usia, dan posisi pemain.
Framework / Metode	Menggunakan XGBoost dengan segmentasi berdasarkan usia dan posisi pemain.
Data yang digunakan	Data dari Sofifa.com yang berisi sekitar 19.000 pemain hingga akhir 2022, dengan 13 atribut utama seperti reputasi internasional, atribut teknis, fisik, kontrak, dan usia.
Temuan	XGBoost menghasilkan akurasi rata-rata sekitar 74%, dengan faktor penentu nilai yang berbeda berdasarkan posisi pemain.

Pembahasan	Penelitian ini menunjukkan bahwa karakteristik pemain dapat mempengaruhi nilai pasar secara berbeda tergantung usia dan posisi. Segmentasi pemain menjadi penting dalam memahami performa model.
Gap Penelitian	Penelitian tersebut hanya berfokus pada XGBoost dan segmentasi karier, sedangkan penelitian ini membandingkan XGBoost dengan Random Forest Regression. Penelitian ini juga menggunakan FIFA 24 Player Stats Dataset dan mengevaluasi performa berdasarkan metrik MAE, RMSE, dan R^2.
Relevansi terhadap penelitian ini	Relevan karena menggunakan XGBoost untuk valuasi pemain sepak bola dan menunjukkan pentingnya analisis berdasarkan karakteristik pemain.
Penelitian Terdahulu 8	
Komponen	**Isi**
Judul	Determinants of Football Players' Valuation: A Systematic Review
Penulis	Franceschi et al.
Sumber jurnal	Journal of Economic Surveys, Vol. 38, No. 3, pp. 577–600.
Tahun	2025
Permasalahan	Penelitian ini membahas faktor-faktor yang menentukan valuasi pemain sepak bola berdasarkan literatur terdahulu selama beberapa dekade.
Framework / Metode	Systematic Literature Review terhadap penelitian valuasi pemain sepak bola.
Data yang digunakan	Literatur penelitian terdahulu mengenai market value dan transfer fee pemain sepak bola selama 30 tahun.
Temuan	Faktor utama dalam penilaian pemain meliputi usia, performa, menit bermain, status tim nasional, kontrak, popularitas, dan karakteristik pemain. Penelitian ini juga menegaskan perlunya metode non-linear seperti Random Forest dan XGBoost karena penelitian sebelumnya banyak menggunakan OLS.
Pembahasan	Penelitian ini memberikan dasar teoritis mengenai faktor yang mempengaruhi nilai pemain dan menunjukkan arah pengembangan metode prediksi yang lebih kuat melalui pendekatan machine learning non-linear.
Gap Penelitian	Penelitian tersebut merupakan systematic review, bukan penelitian eksperimen prediktif. Oleh karena itu, penelitian ini mengisi gap dengan melakukan eksperimen langsung menggunakan Random Forest Regression dan XGBoost Regression pada FIFA 24 Player Stats Dataset.
Relevansi terhadap penelitian ini	Relevan sebagai dasar teori bahwa nilai pasar pemain dipengaruhi oleh banyak faktor dan model non-linear diperlukan untuk menangkap hubungan yang kompleks dalam valuasi pemain.

Penelitian Terdahulu 9	
Komponen	**Isi**
Judul	Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques
Penulis	Al-Asadi & Taşdemir
Sumber jurnal	IEEE Access, Vol. 10, pp. 22631–22645, IEEE.
Tahun	2022
Permasalahan	Penelitian ini membahas prediksi nilai pemain sepak bola menggunakan data FIFA video game dan teknik machine learning.
Framework / Metode	Menggunakan Linear Regression, Multiple Linear Regression, Decision Tree, dan Random Forest Regression.
Data yang digunakan	Dataset FIFA 20 dengan atribut umur, tinggi badan, potensi, reputasi internasional, kaki tidak dominan, posisi, dan passing.
Temuan	Random Forest menjadi model terbaik dengan nilai akurasi sebesar 0,95 atau 95%. Hal ini menunjukkan bahwa model non-linear cocok untuk memprediksi nilai pasar pemain.
Pembahasan	Penelitian ini memperkuat penggunaan FIFA dataset sebagai sumber data dalam prediksi nilai pemain dan menunjukkan keunggulan Random Forest dibandingkan model linear.
Gap Penelitian	Penelitian tersebut belum membandingkan Random Forest dengan XGBoost. Selain itu, dataset yang digunakan masih FIFA 20, sedangkan penelitian ini menggunakan FIFA 24 Player Stats Dataset dan mengevaluasi hasil pada skala logaritmik serta skala nilai pasar asli.
Relevansi terhadap penelitian ini	Relevan karena sama-sama menggunakan dataset FIFA dan Random Forest Regression untuk memprediksi nilai pemain sepak bola.
Penelitian Terdahulu 10	
Komponen	**Isi**
Judul	Machine Learning-Driven Market Value Prediction for European Football Players
Penulis	Tamin et al.
Sumber jurnal	Journal of Computational Mathematics and Data Science, Vol. 15, Article No. 100118, Elsevier.
Tahun	2025
Permasalahan	Penelitian ini membahas prediksi nilai pasar pemain sepak bola Eropa menggunakan pendekatan machine learning.
Framework / Metode	Menggunakan Multiple Linear Regression, Ridge Regression, SVR, dan Random Forest Regression.
Data yang digunakan	Dataset FIFA 22 dari Sofifa.com. Data awal berjumlah 16.710 pemain dan menjadi 12.915 pemain setelah proses pembersihan data.

Temuan	Random Forest Regression menjadi model terbaik dengan nilai error terkecil dan R^2 mencapai 0,96. Model non-linear terbukti lebih unggul dibandingkan model linear dalam memprediksi nilai pasar pemain.
Pembahasan	Penelitian ini menunjukkan bahwa Random Forest Regression memiliki performa tinggi dalam memprediksi nilai pasar pemain sepak bola berbasis dataset FIFA/Sofifa.
Gap Penelitian	Penelitian tersebut belum membandingkan Random Forest Regression dengan XGBoost Regression. Selain itu, penelitian tersebut menggunakan FIFA 22, sedangkan penelitian ini menggunakan FIFA 24 Player Stats Dataset. Penelitian ini juga menambahkan evaluasi pada skala logaritmik dan skala nilai pasar asli untuk kebutuhan interpretasi bisnis.
Relevansi terhadap penelitian ini	Relevan karena sama-sama membahas prediksi nilai pasar pemain sepak bola menggunakan machine learning dan menunjukkan bahwa Random Forest cocok digunakan untuk data tabular pemain.

Berdasarkan Tabel 2.1, penelitian-penelitian terdahulu mengenai prediksi nilai pasar pemain sepak bola telah semakin banyak seiring meningkatnya kebutuhan klub dalam pengambilan keputusan tidak hanya secara finansial maupun teknis seperti proses rekrutmen secara objektif. Penilaian terhadap pemain sepak bola tidak hanya dipengaruhi oleh aspek teknis saja di lapangan, tetapi juga aspek lainnya seperti usia, peforma, menit bermain, popularitas, rating dll [6]. Dalam literatur, penilaian pemain umumnya dibedakan menjadi *market value* dan *transfer fee. market value* sendiri merujuk kepada nilai estimasi pemain secara ekonomi yang diukur dari banyak aspek, sedangkan *transfer fee* merupakan biaya akhir dalam transaksi transfer pemain [6]. Perbedaan inilah yang penting karena penelitian ini berfokus pada prediksi nilai pasar pemain sebagai dara pendukung keputusan rekrutmen.

Beberapa penelitian terdahulu menunjukan bahwa dataset FIFA atau Sofifa dapat dimanfaatkan sebagai sumber data dalam memprediksi nilai pasar pemain. Meskipun beberapa penelitian terdahulu menunjukkan bahwa Support Vector Machine atau Support Vector Regression dapat menghasilkan performa yang baik pada kasus prediksi nilai pemain, penelitian ini tetap menggunakan Random Forest Regression karena karakteristik data yang digunakan berbentuk tabular, memiliki

banyak atribut numerik, memiliki potensi hubungan non-linear, serta memuat fitur dengan variasi nilai yang cukup besar. Random Forest juga relevan karena beberapa penelitian berbasis FIFA/Sofifa menunjukkan bahwa model berbasis pohon mampu menghasilkan performa tinggi dalam prediksi nilai pemain. Dengan demikian, pemilihan Random Forest dalam penelitian ini bukan untuk mengabaikan performa SVM, tetapi untuk menguji model ensemble berbasis tree yang lebih sesuai dengan kebutuhan interpretasi data tabular dan perbandingan dengan XGBoost. Jurnal yang ditulis Behravan dan Razavi menggunakan dataset FIFA 20 dengan metode APSO-clustering dan PSO-SVR untuk memprediksi nilai pemain [15]. Peneliti lain seperti Al-Asadi dan tasdemir juga memanfaatkan dataset FIFA 20 dan menemukan bahwa model Random Forest Regression memperoleh peforma terbaiknya dengan nilai R^2 sebesar 0,95 [16]. Temuan serupa ditunjukkan oleh Tamim et al. yang menggunakan dataset FIFA 22, di mana Random Forest Regression mencapai R^2 sebesar 0,96 hingga 0,99 [17]. Melalui hasil yang didapat hubungan antara atribut pemain dan nilai pasar cenderung bersifat non-linear, sehingga model berbasis pohon keputusan lebih sesuai dibandingkan model linear tradisional.

Selain FIFA, beberapa penelitian menggunakan dataset dunia nyata seperti *transfermarket*, FBRef, SciSports [18]. Li et al. dalam jurnalnya membandingkan Decision Tree, Random Forest, SVR, dan XGBoost menggunakan data Transfermarkt, FBRef, dan Sofifa, dimana hasil prediksi menunjukan XGBoost menjadi model terbaik dengan RMSE sebesar 0,717 dan R^2 sebesar 0,695 [19]. van Arem et al. menunjukkan bahwa Random Forest lebih sesuai untuk prediksi nilai pemain, sedangkan XGBoost unggul dalam prediksi kualitas pemain [20]. Hasil ini memperlihatkan bahwa baik Random Forest maupun XGBoost memiliki potensi kuat dalam menangani data nilai pemain yang kompleks.

Perkembangan penelitian terbaru juga menunjukkan perhatian terhadap interpretabilitas, segmentasi pemain, dan prediksi nilai masa depan. Shen menggunakan Random Forest Regression, XGBoost, CatBoost, SVR, serta model ansambel berbasis Dempster-Shafer Theory untuk memprediksi nilai pemain menggunakan data FIFA 19 dan data statistik dunia nyata [21]. Penelitian tersebut juga menggunakan SHAP dan FAST untuk memahami pengaruh fitur terhadap nilai

pemain. Khalife et al. menggunakan Gradient Boosting atau XGBoost dengan segmentasi berdasarkan usia dan posisi pemain, serta menunjukkan bahwa penambahan fitur potensi masa depan meningkatkan akurasi rata-rata sebesar 74% [22]. Temuan ini menunjukkan bahwa faktor penentu nilai pemain dapat berbeda berdasarkan posisi, usia, dan tahap karier.

Berdasarkan penelitian terdahulu, dapat disimpulkan bahwa prediksi nilai pasar pemain sepak bola merupakan permasalahan regresi yang kompleks karena dipengaruhi oleh faktor teknis, fisik, performa, usia, reputasi, kontrak, posisi, dan konteks pasar. Sebagian besar penelitian menunjukkan bahwa model non-linear seperti Random Forest dan XGBoost lebih mampu menangkap pola hubungan antarvariabel dibandingkan model linear. Namun, penelitian sebelumnya masih memiliki keterbatasan dari sisi dataset, target prediksi, dan fokus perbandingan model. Beberapa penelitian menggunakan dataset FIFA versi lama, seperti FIFA 19, FIFA 20, dan FIFA 22. Jurnal lainnya menggunakan data Transfermarkt, FBRef, atau data eksklusif seperti SciSports. Selain itu, tidak semua penelitian membandingkan Random Forest Regression dan XGBoost Regression secara langsung sebagai fokus utama.

Secara keseluruhan, penelitian ini diarahkan untuk membandingkan kinerja Random Forest Regression dan XGBoost Regression dalam memprediksi nilai pasar pemain sepak bola menggunakan FIFA 24 Player Stats Dataset. Penggunaan dataset FIFA 24 memberikan pembaruan dari sisi data, sedangkan perbandingan kedua model memberikan dasar untuk menentukan algoritma yang lebih sesuai dalam konteks prediksi nilai pasar pemain. Evaluasi dilakukan menggunakan MAE, RMSE, dan R^2 agar hasil perbandingan dapat dianalisis secara objektif. Dengan demikian, penelitian ini diharapkan dapat mendukung pengembangan sistem pendukung keputusan rekrutmen pemain berbasis data.

2.2 Teori yang berkaitan

2.2.1 Nilai Pasar dan Rekrutmen Berbasis Data

Industri sepak bola saat ini berkembang yang melibatkan aspek olahraga, bisnis dan analisis data. Pemain tidak hanya dipandang sebagai alat atau bagian dari kebutuhan taktis tim, tetapi juga sebagai aset organisasi yang memiliki nilai

olahraga dan komersial [3] [23] [24]. Oleh karena itu, keputusan terkait proses rekrutmen pemain menjadi salah satu keputusan strategis bagi klub karena melibatkan tidak hanya peforma tim tapi juga efisiensi terhadap anggaran serta keberlanjutan finansial klub [3].

Proses rekrutmen pemain menjadi bagian keputusan investasi klub [23]. Klub perlu melakukan penilaian apakah biaya yang dikeluarkan oleh tim finansial untuk memperoleh seorang pemain sebanding dengan manfaat kebutuhan tim dalam jangka pendek maupun panjang sesuai dengan bagaimana prioritas perekrutan, baik dalam bentuk peforma tim, daya jual nama tim, maupun potensi penjualan pemain tersebut [3] [23] [25] [26]. Dengan demikian proses rekrutmen tidak cukup hanya mengandalkan intuisi, pengalaman atau hasil peniliain subjektif, tetapi perlu didukung oleh data aktual yang memberikan gambaran penilaian pemain secara terukur.

Nilai pasar pemain menjadi atribut penting dalam konsep proses rekrutmen berbasis data [24] [25] [27]. Nilai pasar merupakan estimasi dari nilai bisnis atau ekonomi yang dimiliki seorang pemain pada periode waktu tertentu yang dipengaruhi oleh faktor usia, peforma, potensi, serta dinamika pasar [28]. Secara konseptual, faktor yang memengaruhi value pemain dapat dikelompokkan ke dalam beberapa aspek, yaitu aspek performa, karakteristik individu, kontraktual, reputasi, dan pasar. Aspek performa berkaitan dengan kontribusi teknis, konsistensi permainan, produktivitas, dan kemampuan pemain dalam menjalankan perannya. Aspek karakteristik individu mencakup usia, posisi, kondisi fisik, dan potensi perkembangan. Aspek kontraktual berkaitan dengan durasi kontrak dan posisi tawar klub dalam proses negosiasi. Aspek reputasi dan popularitas berkaitan dengan citra pemain, status tim nasional, dan nilai komersial. Sementara itu, aspek pasar mencakup kebutuhan klub terhadap posisi tertentu, kelangkaan pemain, permintaan pasar, serta pembanding dari data historis transfer pemain sejenis [22], [23], [24].

Nilai pasar pemain berbeda dengan biaya transfer, dimana biaya transfer merupakan harga transaksi akhir yang disepakati antar klub. Sedangkan nilai pasar merupakan estimasi nilai pemain sebelum transaksi terjadi. Hal ini penting karena

biaya transfer dipengaruhi oleh kondisi yang berbeda seperti negosiasi dan durasi kontrak pemain.

Dalam ranah sistem Informasi, data pemain diolah menjadi informasi pendukung dalam pengambilan keputusan proses rekrutmen pemain [3]. Atribut seperti usia, kemampuan fisik, rating, potensi dan peforma lainnya dapat digunakan sebagai input untuk membangun model prediksi nilai pasar [28]. Informasi yang dihasilkan bukan semata merubah keputusan sepenuhnya untuk tim finansial maupun tim rekrutmen pemain melainkan menjadi jembatan atau alat bantu dalam menetapkan prioritas transfer pemain secara objektif [27] [29] [30].

Penelitian ini memposisikan nilai pasar pemain sebagai variabel dependen atau target yang akan diprediksi nilainya menggunakan model regresi Random Forest dan XGBoost. Pemilihan nilai pasar sebagai target prediksi karena nilai tersebut dapat menjadi dasar awal untuk tim finansial menilai kewajaran ekonomi dalam proses rekrutmen serta sesuai dengan kebutuhan tim rekrutmen [3] [23]. Dengan adanya model prediksi berbasis data, klub terutama tim finansial dan rekrutmen dapat memperoleh informasi pendukung guna membandingkan estimasi nilai pemain berdasarkan atribut yang dimiliki dalam dataset. Dengan demikian, prediksi nilai pasar pemain dapat diposisikan sebagai sistem pendukung keputusan rekrutmen berbasis data [3].

2.2.2 Sistem Pendukung Keputusan dalam Rekrutmen Pemain

Sistem Pendukung Keputusan merupakan konsep yang digunakan untuk membantu proses pengambilan keputusan dengan melalui pemanfaatan data, model dan analisis [29] [30]. Sistem pendukung keputusan ini tidak dimaksudkan untuk menggantikan peran manusia dalam pengambilan keputusan. Justru sistem ini menyediakan informasi yang dapat memperkuat kualitas keputusan. Sistem ini berfungsi sebagai pendukung pengambilan keputusan dalam kasus yang kompleks yaitu kasus dimana penetapan keputusannya tidak ada aturan tetap, tetapi membutuhkan pertimbangan tergantung sebuah kondisi.

Umumnya sistem pendukung keputusan terdiri dari komponen data, model dan penyajian informasi. Komponen data berfungsi untuk menyediakan informasi yang diperlukan dalam proses analisis, komponen model digunakan untuk proses

pengolahan data agar menghasilkan informasi. Kemudian informasi tersebut menjadi bentuk yang mudah dipahami oleh aktor yang mengambil keputusan. Dalam penelitian ini, peran atribut sebagai komponen data yang menyediakan informasi, sedangkan Random Forest dan XGBoost Regression yang akan mengolah dan memproses data tersebut agar menjadi sebuah informasi yang digunakan dalam mendukung keputusan transfer pemain.

Dalam proses rekrutmen pemain sepak bola, kebutuhan terhadap sistem pendukung keputusan menjadi penting karena klub harus mengevaluasi banyak kandidat pemain dengan karakteristik yang berbeda-beda, karena proses rekrutmen tidak hanya berkaitan dengan kualitas teknis pemain, tetapi juga kebutuhan posisi, anggaran, potensi perkembangan, risiko investasi, nilai pasar, dan kemungkinan nilai jual kembali. Dengan demikian, pendekatan berbasis data dapat membantu proses evaluasi pemain menjadi lebih terukur dan sistematis.

Secara keseluruhan, penelitian ini tidak hanya berfokus pada pembangunan model machine learning, tetapi juga pada pemanfaatan hasil prediksi sebagai informasi pendukung dalam pengambilan keputusan. Hubungan antara atribut yang digunakan sebagai informasi untuk mengolah data dengan model prediksi yang digunakan, teori sistem pendukung keputusan menjadi landasan penting untuk menghubungkan perbandingan kinerja XGBoost Regression dan Random Forest Regression dengan tujuan praktis penelitian yaitu mendukung proses rekrutmen pemain secara lebih terukur dan berbasis data.

2.2.3 Machine Learning, Regresi, dan Ensemble Learning

Machine Learning merupakan cabang ilmu kecerdasan buatan yang dimana model akan mempelajari sebuah pola dari data dan menggunakan pola tersebut untuk menghasilkan prediksi atau kesimpulan terhadap data baru [31] [32] [33]. Berbeda dengan kecerdasan buatan atau AI yang semua prosesnya bergantung pada pengguna sedangkan model machine learning dapat melakukan pekerjaan semi-otomatis berdasarkan data pelatihan yang digunakan sebagai metode belajar. Sehingga machine learning relevan digunakan terhadap permasalahan yang memiliki pola hubungan yang kompleks.

Dalam penelitian ini, pendekatan yang digunakan termasuk dalam s*upervised learning* karena model dilatih atau input data menggunakan data masa lalu untuk dipelajari untuk menarik kesimpulan kasus serupa kedepannya yaitu disebut juga target variabel [31] [34] . Input berupa atribut pemain, sedangkan target output berupa nilai pasar pemain. Karena target yang diprediksi berbentuk nilai numerik kontinu, maka dari itu jenis tugas prediksi yang dilakukan yaitu regresi.

Prediksi nilai pasar pemain sepak bola merupakan permasalahan yang kompleks, karena nilai pemain dipengaruhi banyak faktor seperti usia, posisi, performa, potensi, reputasi, kemampuan teknis, dan atribut fisik. Selain itu, hubungan antarvariabel tersebut tidak selalu bersifat linear, sehingga model sederhana berisiko tidak mampu menangkap pola data secara optimal. Oleh sebab itu, diperlukan pendekatan model yang lebih fleksibel untuk mempelajari hubungan non-linear antar atribut.

Salah satu pendekatan yang cocok dalam menangani permasalahan tersebut adalah *ensemble learning*. *Ensemble learning* adalah salah satu pendekatan machine learning yang menggabungkan beberapa model pembelajar untuk menghasilkan prediksi yang lebih stabil dan memiliki kemampuan generalisasi lebih baik dibandingkan model tunggal [35] [36]. Pendekatan ini banyak digunakan pada data tabular karena mampu menangani banyak fitur, pola non-linear, dan interaksi antarvariabel.

Metode memiliki dua pendekatan utama, yaitu *bagging* dan *boosting. Bagging* memetakan model ke bentuk beberapa model secara paralel menggunakan variasi data, kemudian menggabungkan hasil prediksinya untuk menghasilkan hasil akhir yang lebih stabil dimana Random Forest Regression termasuk dalam pendekatan ini. Sedangkan *boosting* membangun model secara bertahap, di mana model berikutnya berfokus memperbaiki kesalahan model sebelumnya dimana model yang menggunakan pendekatan ini yaitu XGBoost Regression [37].

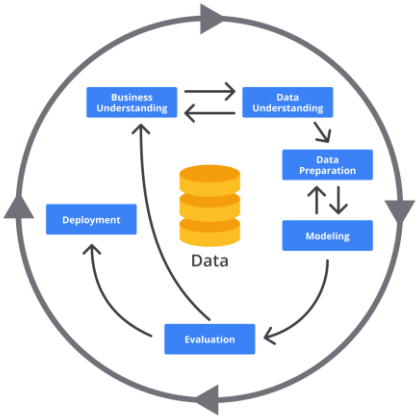
Secara keseluruhan, machine learning, regresi, dan ensemble learning menjadi landasan teoretis yang saling berkaitan dalam penelitian ini. Dimana Machine learning digunakan sebagai pendekatan utama untuk mempelajari pola dari data, dan pemilihan regresi digunakan karena target penelitian berupa nilai numerik

kontinu, sedangkan *ensemble learning* menjadi dasar penggunaan Random Forest Regression dan XGBoost Regression. Kedua algoritma yang digunakan sangat relevan untuk dibandingkan dalam memprediksi nilai pasar pemain sepak bola sebagai informasi pendukung dalam proses transfer pemain.

2.3 Framework/Algoritma yang digunakan

2.3.1 *Framework* CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) merupakan kerangka kerja yang banyak digunakan untuk mengelola proyek data mining dan machine learning secara sistematis. Kerangka ini terdiri atas enam tahapan, yaitu *business understanding, data understanding, data preparation, modeling*, *evaluation* dan *deployment.* Semua tahapan ini saling berkaitan dan berjalan secara terstruktur, dimana tahapan – tahapan ini menghubungkan dari mulai kebutuhan masalah, kualitas data, transformasi data, proses pemodelan, hingga pemanfaatan hasil analisis [31]. Tahapan yang ada pada CRISP-DM ini digabarkan pada gambar 2.1.



Gambar 2. 1 CRISP-DM FRAMEWORK

sumber: [38]

Berdasarkan gambar 2.1, tahap pertama yang dilakukan yaitu *business understanding.* Pada tahapan ini berfokus pada pemahaman masalah dan tujuan dari penelitian. Dalam penelitian Perbandingan Kinerja XGBoost Regression dan Random Forest Regression dalam prediksi nilai pasar pemain sepak bola sebagai

pendukung keputusan rekrutmen adalah kebutuhan terhadap pendekatan berbasis data dalam memprediksi nilai pasar pemain sepak bola sebagai pendukung pengambilan keputusan secara objektif. Tidak hanya itu, penelitian ini juga membandingkan kinerja dari dua algoritma yang digunakan yaitu Random Forest Regression dan XGBoost Regression untuk menetapkan model non linear mana yang lebih sesuai digunakan dalam kasus [31] [32].

Tahapan kedua yaitu *data understanding*. Tahap *data understanding* merupakan tahapan untuk menganalisa dan memahami struktur, jumlah, atribut, tipe, kualitas dan kondisi data. Pada dasarnya tahapan ini bertujuan untuk memahami secara deskriptif apa yang terjadi pada data, sebagai gambaran untuk melakukan analisis mendalam [32] [33].

Tahapan ketiga adalah *data preparation* atau *preprocessing*. Pada tahapan ini data hasil dari tahapan *data understanding* akan diproses dalam proses persiapan data agar data tersebut dapat digunakan ketika membangun sebuah model *machine learning*. Tahapan yang termasuk dalam data preparation yaitu meliputi pembersihan data, penghapusan atribut yang tidak relevan, penanganan missing value, penghapusan data duplikat, konversi data dari kategori menjadi numerik, transformasi data, serta berbagai tahapan preparasi pada data lainnya hingga data siap dibagi menjadi data *training* dan *test*. Tahapan ini penting karena akurasi dari prediksi model machine learning yang dibangun bergantung pada kualitas datanya [33] [39].

Tahapan keempat yaitu *modeling.* Pada tahapan ini, data yang telah siap melalui proses sebelumnya yaitu preparasi akan digunakan untuk membangun model prediksi. Model akan membangun data dengan melakukan *training* sesuai jumlah data yang telah dibagi pada tahapan preparasi untuk melakukan pembelajaran dan setelah itu melakukan pengujian [31] [32].

Tahapan kelima yaitu *evaluation.* Tahapan ini dilakukan untuk mengukur atau menguji kinerja dari akurasi model *machine learning* yang dibangun. Evaluasi dilakukan dengan membandingkan nilai aktual dan nilai prediksi menggunakan metrik atau alat ukur seperti *Mean Absolute Error, Root Mean Squared Error* dan *R-Squared* untuk kasus regresi [31] - [34].

Tahapan terakhir yaitu *deployment.* Tahapan ini merepresentasikan pemanfaatan dari model yang telah dibangun. Dalam konteks penelitian Perbandingan Kinerja XGBoost Regression dan Random Forest Regression dalam prediksi nilai pasar pemain sepak bola sebagai pendukung keputusan rekrutmen bukan implementasi sistem secara penuh ke lingkungan produksi. Deployment dapat dimaknai sebagai hasil penyajian model, hasil evaluasi, visualisasi serta rekomendasi model yang optimal dalam mendukung informasi pendukung keputusan [30] - [33].

Secara keseluruhan pada tahapan CRISP-DM, kerangka kerja ini digunakan karena sesuai dengan karakter penelitian prediktif berbasis data. Kerangka kerja ini membantu memastikan bahwa penelitian tidak hanya berfokus pada algoritma, tetapi juga memperhatikan keterkaitan antara masalah penelitian, kualitas data, pembangunan model, proses evaluasi kinerja, dan pemanfaatan hasil analisis sesuai tahapan pada gambar 2.1.

2.3.2 Pengolahan Data dan Atribut Kategorikal dalam Pemodelan

Seperti yang telah dijelaskan pada tahapan *data preparation* sebelum model machine learning dibangun dalam konteks penelitian ini yaitu Random Forest Regression dan XGBoost Regression, data perlu disiapkan agar sesuai dengan kebutuhan model. Salah satu aspek penting dalam membangun model yaitu bagaimana pengolahan pada variabel atau atribut dari data. Apabila variabel target yang digunakan untuk tahapan regresi berupa kategorikal bukan numerik, Oleh karena itu target variabel perlu dikonversi ke dalam bentuk numerik agar dapat digunakan ke dalam model regresi [33].

Ketika data yang telah ditransformasi ke dalam bentuk numerik ternyata distribusinya sangat condong ke kanan, Salah satu metode yang bisa digunakan yaitu melakukan transformasi logaritmik. Transformasi logaritmik digunakan untuk mengurangi pengaruh nilai ekstrem dan membuat distribusi dari data target menjadi lebih seimbang. Dalam kasus penelitian ini, data target asli akan direpresentasikan sebagai variabel *value_numeric*, sedangkan target hasil dari transformasi logaritmik yaitu *value_log*. Hal ini dilakukan untuk menjaga distribusi dari model yang diukur

lebih stabil, sehingga *value_numeric* tetap digunakan untuk kebutuhan hasil prediksi dalam skala nilai asli [35] [39].

Tahapan lainnya untuk memproses atribut data yang berupa teks yaitu dengan menggunakan metode *Frequency encoding.* Metode ini bekerja dengan mengganti setiap kategori pada atribut tersebut ke bentuk frekuensi kemunculannya di dalam dataset. Dengan teknik ini, atribut yang nilainya kategori tidak lagi di representasikan berupa teks, melainkan sebagai angka berdasarkan jumlah kemunculannya [36]. Teknik ini cocok untuk atribut dengan jumlah kategori tinggi karena mampu mengubah data kategorikal menjadi numerik tanpa menambah dimensi fitur secara berlebihan.

Pengolahan target dan atribut kategorikal menjadi bagian penting karena kedua model yang digunakan membutuhkan data dalam bentuk numerik. Dengan melakukan proses ini data yang akan digunakan nilainya lebih sesuai dalam tahapan pemodelan penelitian ini yaitu Random Forest Regression dan XGBoost Regression.

2.3.3 Random Forest Regression

Random Forest Regression adalah algoritma machine learning berbasis *ensemble learning* yang digunakan untuk menyelesaikan permasalahan regresi. Algoritma ini sebagai solusi atau peningkatan dari model *decision tree* dimana metode ini bekerja dengan membangun banyak *decision tree,* kemudian menggabungkan hasil prediksi dari seluruh pohon untuk menghasilkan prediksi akhir. Pada kasus regresi, hasil diperoleh melalu penggabungan rata-rata prediksi dari seluruh pohon [34] - [39]. Pendekatan ini membuat Random Forest Regression lebih stabil dibandingkan *decision tree* tunggal karena keputusan akhir tidak hanya bergantung pada satu model.

Random Forest menggunakan pendekatan *bagging* atau *bootstrap aggregating*. Yaitu, setiap pohon dibangun dari sampel data yang berbeda, sehingga setiap pohon dapat mempelajari variasi pola yang berbeda pula. Selain itu, Random Forest juga menggunakan pemilihan fitur secara acak pada proses pembentukan pohon. Kombinasi antara *bootstrap sampling* dan *random feature selection* bertujuan untuk

mengurangi korelasi antar pohon, menurunkan risiko *overfitting*, serta meningkatkan kemampuan generalisasi model terhadap data baru [31].

Dalam konteks prediksi nilai pasar pemain sepak bola, Random Forest Regression menjadi model yang relevan karena mampu menangani data tabular yang memiliki banyak atribut. Nilai pasar pemain dipengaruhi oleh banyak faktor, seperti usia, posisi, performa, kemampuan teknis, potensi, reputasi, dan karakteristik lainnya. Hubungan antara faktor-faktor tersebut dengan nilai pasar tidak selalu linear. Oleh karena itu, algoritma berbasis pohon seperti Random Forest Regression dapat menjadi pilihan yang tepat karena mampu menangkap hubungan non-linier dan interaksi antarvariabel.

Kelebihan utama Random Forest Regression adalah cenderung stabil dalam memprediksi nilai target dan ketahanannya terhadap *overfitting* dibandingkan *decision tree* tunggal. Karena model ini menggabungkan banyak pohon, kesalahan dari satu pohon dapat dikurangi oleh kontribusi pohon lainnya. Selain itu, Random Forest juga tidak terlalu sensitif terhadap pengaturan hiperparameter tertentu dibandingkan beberapa model ensemble lainnya, dan juga dapat memberikan informasi mengenai tingkat kepentingan fitur, sehingga peneliti dapat memahami variabel mana yang memiliki kontribusi relatif besar terhadap prediksi nilai pasar pemain [31] [35].

Meskipun demikian, Random Forest Regression juga memiliki keterbatasan atau kelemahan. Model ini dapat membutuhkan sumber daya komputasi yang lebih besar apabila jumlah pohon dan jumlah fitur yang digunakan cukup banyak. Selain itu, hasil dan visualisasi model Random Forest tidak sesederhana model linear karena keputusan akhir berasal dari gabungan banyak pohon. Oleh karena itu, penggunaan Random Forest Regression perlu disertai evaluasi model yang cermat untuk memastikan performa yang diperoleh tidak hanya bagus pada data latihnya tetapi ketika melakukan prediksi pada data uji.

Dalam penelitian ini, Random Forest Regression digunakan sebagai salah satu model utama yang dibandingkan dengan XGBoost Regression. Pemilihan model ini didasarkan pada kemampuannya dalam menangani data nonlinier dan sebagai salah satu metode *ensemble learning*. Dengan karakteristik yang dimilikinya,

Random Forest menjadi model pembanding yang relevan untuk menilai kinerja XGBoost Regression dalam memprediksi nilai pasar pemain sepak bola.

2.3.4 XGBoost Regression

XGBoost Regression adalah algoritma machine learning berbasis *gradient boosting* yang digunakan untuk menyelesaikan permasalahan regresi. Berbeda dengan Random Forest Regression yang membangun banyak pohon secara sekaligus, XGBoost membangun pohon secara bertahap. Setiap pohon baru dibuat untuk memperbaiki kesalahan prediksi dari model sebelumnya. Dengan mekanisme tersebut, XGBoost berusaha meningkatkan akurasi model secara iteratif melalui proses pembelajaran yang berurutan [31] [35].

XGBoost adalah pengembangan dari metode boosting yang dimaksudkan meningkatkan efisiensi dalam menangani berbagai jenis data kompleks. Salah satu karakteristik penting XGBoost adalah adanya mekanisme regularisasi yang digunakan untuk mengendalikan kompleksitas model. Regularisasi ini membantu mengurangi risiko *overfitting*, terutama ketika model dilatih pada data dengan banyak atribut atau pola yang kompleks. Selain itu, XGBoost juga menyediakan berbagai parameter yang dapat disesuaikan, seperti *learning rate* yaitu jumlah seberapa cepat model belajar kesalahan , jumlah pohon, kedalaman pohon, sub-sampling, serta parameter regularisasi [37].

XGBoost Regression digunakan untuk memprediksi tipe data numerik kontinu. Hal ini sesuai dengan tujuan penelitian karena nilai pasar pemain sepak bola yang berbentuk angka. XGBoost relevan digunakan dalam penelitian ini karena nilai pasar pemain dipengaruhi oleh banyak variabel yang saling berinteraksi dan tidak selalu memiliki hubungan linear. Melalui pendekatan boosting, XGBoost memiliki kemampuan untuk memperbaiki kesalahan prediksi secara bertahap sehingga berpotensi menghasilkan model yang akurat pada data tabular.

Kelebihan utama XGBoost Regression adalah kemampuannya dalam menangkap pola kompleks, menangani hubungan nonlinier, dan memberikan performa yang tinggi pada banyak permasalahan prediktif. Algoritma ini juga dikenal efisien secara komputasi dan memiliki fleksibilitas dalam pengaturan parameter. Dalam penelitian nilai pasar pemain, keunggulan tersebut penting

karena data pemain dapat memiliki variasi yang tinggi, baik dari sisi karakteristik individu maupun atribut performa [39] [40].

Namun, XGBoost Regression juga memiliki beberapa keterbatasan. Model ini cenderung lebih sensitif terhadap pemilihan hiperparameter dibandingkan Random Forest Regression. Pengaturan parameter yang kurang tepat dapat menyebabkan model terlalu sederhana atau justru terlalu kompleks. Selain itu, karena proses pembelajaran dilakukan secara bertahap, waktu pelatihan dapat meningkat apabila jumlah pohon dan kedalaman model terlalu besar. Oleh karena itu, penggunaan XGBoost perlu disertai strategi evaluasi yang adil dan konsisten agar performanya dapat dibandingkan secara objektif dengan model lain.

Dalam penelitian ini, XGBoost Regression dipilih sebagai model pembanding utama karena mewakili pendekatan boosting dalam *ensemble learning*. Perbandingan dengan Random Forest Regression menjadi penting karena kedua algoritma sama-sama berbasis pohon keputusan, tetapi memiliki mekanisme pembelajaran yang berbeda. Random Forest membangun banyak pohon secara bercabang dan menggabungkan prediksinya, sedangkan XGBoost membangun model secara bertahap untuk memperbaiki kesalahan sebelumnya. Perbedaan tersebut menjadi dasar penelitian untuk menguji model mana yang lebih sesuai dalam memprediksi nilai pasar pemain sepak bola.

2.3.5 Evaluasi Kinerja Model Regresi

Tahapan evaluasi kinerja model regresi merupakan tahapan penting untuk menilai kualitas prediksi yang dihasilkan oleh model. Dalam penelitian prediktif, model tidak cukup dinilai berdasarkan kemampuannya menyesuaikan data latih, tetapi harus diuji pada data yang belum pernah digunakan dalam proses pelatihan (data testing). Hal ini dilakukan untuk mengetahui apakah model memiliki kemampuan generalisasi yang baik pada data baru atau hanya menghafal pola pada data pelatihan.

Dalam penelitian ini, evaluasi dilakukan untuk membandingkan performa Random Forest Regression dan XGBoost Regression dalam memprediksi nilai pasar pemain sepak bola. Karena target yang diprediksi berupa nilai numerik kontinu, maka metrik evaluasi yang digunakan adalah metrik regresi. Beberapa

metrik yang umum digunakan dalam evaluasi model regresi adalah Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan koefisien determinasi atau R^2 [34] [35] [39].

Mean Absolute Error atau MAE digunakan untuk mengukur rata-rata besar kesalahan absolut antara nilai aktual dan nilai prediksi. Metrik ini mudah diimplementasikan karena memiliki satuan yang sama dengan target prediksi. MAE dapat menunjukkan rata-rata kesalahan prediksi nilai pasar pemain. Semakin kecil nilai MAE, semakin kecil rata-rata penyimpangan prediksi model dari nilai aktual [34] [35] [39].

Mean Squared Error atau MSE digunakan untuk mengukur rata-rata kuadrat kesalahan prediksi. Karena error dikuadratkan, MSE memberikan penalti yang lebih besar terhadap kesalahan prediksi yang ekstrem. Metrik ini berguna untuk mengetahui apakah model menghasilkan kesalahan besar pada sebagian kondisi. Dalam prediksi nilai pasar pemain, kesalahan besar dapat menjadi perhatian penting karena dapat berdampak pada hasil kewajaran nilai pemain [34] [35] [39].

Root Mean Squared Error atau RMSE merupakan akar dari MSE. RMSE memiliki kelebihan karena kembali ke satuan asli target, sehingga lebih mudah diimplementasikan dibandingkan MSE. RMSE tetap sensitif terhadap kesalahan besar, sehingga dapat digunakan untuk menilai apakah model memiliki penyimpangan nilai prediksi yang besar pada pemain tertentu. Semakin kecil nilai RMSE, semakin baik kemampuan model dalam menghasilkan prediksi yang mendekati nilai aktual [34] [35] [39].

Koefisien determinasi atau R^2 digunakan untuk mengetahui seberapa besar variasi nilai target yang dapat dijelaskan oleh model. Nilai R^2 yang semakin mendekati 1 menunjukkan bahwa model mampu menjelaskan variasi nilai pasar dengan lebih baik. Dalam penelitian ini, R^2 digunakan untuk melengkapi metrik berbasis error karena MAE, MSE, dan RMSE menunjukkan besar kesalahan prediksi, sedangkan R^2 menunjukkan kemampuan model dalam menjelaskan variasi nilai pasar pemain.

Karena pada tahap preparasi menggunakan proses transformasi logaritmik pada variabel targetnya atau independen. Evaluasi dilakukan pada dua skala yaitu skala

logaritmik dan skala asli. Skala logaritmik dilakukan untuk menilai peforma model terhadap variabel target yang digunakan pada saat pelatihan. Sementara evaluasi skala asli digunakan untuk memahami besar kesalahan prediksi dalam satuan asli yang lebih mudah diinterpretasikan artinya nilai pasar pemain dalam bentuk nominal.

Proses validasi juga dilakukan dengan menggunakan *cross-validation. Cross-validation* adalah metode evaluasi untuk melihat kestabilan peforma model dengan membagi data latih ke dalam beberapa bagian atau disebut juga fold. Melalui pendekatan ini, model prediksi yang diuji dibagi antara data latih dan validasi sehingga hasil evaluasi tidak hanya bergantung pada satu pembagian data tapi juga membantu memberikan gambaran yang lebih stabil terkait kemampuan model dalam memprediksi nilai [40].

Analisis evaluasi residual juga digunakan sebagai tahapan bagaimana selisih antara nilai aktual dan nilai prediksi. Analisis residual digunakan untuk melihat pola kesalahan dalam prediksi model. Apabila nilai residual tersebar secara seimbang dan tidak menunjukan pola tertentu, maka model dapat dianggap memiliki kesalahan prediksi yang lebih stabil. Sebaliknya jika analisis residual menunjukan suatu pola yang jelas, maka dapat diindikasikan model tersebut belum sepenuhnya menangkap karakteristik pada data.

Proses evaluasi juga dilakukan untuk mengukur segmentasi nilai pasar pemain. Segmentasi ini digunakan untuk mengetahui apakah peforma model berbeda pada kelompok pemain bernilai rendah, menengah, tinggi dan elite. Tahapan ini penting karena distribusi pasar pemain umumnya tidak seimbang dimana sebagian besar pemain berada pada nilai pasar rendah hingga menengah, sedangkan pemain bernilai sangat tinggi jumlahnya lebih sedikit. Kondisi seperti itu yang dapat menyebabkan model memprediksi kelompok mayoritas, tetapi sulit untuk kelompok minoritas [34].

Proses evaluasi terakhir bertujuan untuk memperkuat perbandingan kinerja dari model Random Forest Regression dan XGBoos Regression, dilakukannya uji statistik yang dapat digunakan sebagai analisis tambahan. Uji statistik seperti paired t-test dan Wilcoxon dapat digunakan untuk mengetahui apakah perbedaan error

antara kedua model bersifat signifikan. Pengukuran efek seperti Cohen's d juga dapat digunakan untuk melihat besar kecilnya perbedaan kinerja kedua model sehingga perbandingan model tidak hanya didasarkan pada nilai metrik, tetapi juga didukung oleh analisis statistik [34].

Secara keseluruhan tahapan, *framework* dan algoritma yang digunakan dalam penelitian ini tidak hanya berfungsi untuk menghasilkan model prediksi, tetapi juga mendukung proses transformasi data menjadi informasi yang bernilai. Hal ini sesuai dengan fokus penelitian di ranah Big Data Sistem Informasi yaitu pemanfaatan teknologi, data dan model analitis untuk mendukung proses pengambilan keputusan secara objektif dan terukur.

2.4 Tools/software yang digunakan

2.4.1 Python dan Jupyter Notebook

Pada penelitian ini bahasa pemrograman yang digunakan yaitu python. Python adalah bahasa pemrograman interpertasi yang tidak melalui proses compile ketika kode dijalankan. Python dipilih menjadi bahasa pemrograman utama karena bahasanya memiliki ekosistem yang mendukung untuk proses analisis data, *machine learning*, komputasi numerik, dan visualisasi. Dalam penelitian ini python digunakan dari mulai membaca dataset, memeriksa struktur data, membangun model prediksi hingga mengevaluasi model yang dibangun. Penggunaan python juga membantu proses penelitian menjadi efisien karena tahapan analisis dapat dilakukan dalam satu alur kerja yang terintegrasi.

Jupyter Notebook adalah lingkungan kerja utama untuk menjalankan kode python dan mendokumentasikan proses penelitian. Melalui jupyter notebook, *syntax* atau perintah kode untuk mesin dijalankan dalam membentuk tabel, grafik, serta output analisis data dalam satu dokumen ipnyb. Karakteristik tersebut sesuai dengan penelitian berbasis data karena setiap tahapan, mulai dari *import* dataset, pemahaman data, preparasi, pemodelan hingga evaluasi dapat disusun secara terstruktur dan mudah.

2.4.2 Pandas, NumPy, *Scikit-learn*, dan XGBoost

Pandas digunakan sebagai *library* utama untuk mengolah data tabular. Pandas digunakan untuk membaca dataset FIFA 24 Player Stats, menampilkan struktur data, memeriksa jumlah baris dan kolom, mengidentifikasi tipe data, memeriksa *missing value*, menghapus data duplikat, melakukan seleksi atribut, serta menyiapkan dataset sebelum digunakan dalam proses pemodelan. Pandas juga digunakan untuk mengelola kolom target dan atribut fitur agar sesuai dengan kebutuhan model regresi.

NumPy digunakan untuk mendukung proses komputasi numerik dalam penelitian. *Library* ini membantu dalam pengolahan array, perhitungan matematis, transformasi numerik, serta operasi yang berkaitan dengan data dalam bentuk angka. NumPy relevan digunakan karena proses pemodelan machine learning membutuhkan data numerik yang konsisten dan dapat diproses secara efisien.

Scikit-learn digunakan sebagai *library* machine learning yang mendukung beberapa proses utama dalam penelitian. Scikit-learn digunakan untuk melakukan pembagian data latih dan data uji, membangun model Random Forest Regression, melakukan *cross-validation*, serta menghitung metrik evaluasi model regresi. *library* ini dipilih karena menyediakan fungsi machine learning yang terstruktur dan umum digunakan dalam penelitian prediktif berbasis data tabular. Selain itu, Scikit-learn juga digunakan untuk menghitung metrik evaluasi seperti *Mean Absolute Error, Root Mean Squared Error,* dan *R-Squared*. Oleh karena itu, Scikit-learn berperan penting dalam menjaga konsistensi proses eksperimen dan evaluasi model.

XGBoost digunakan sebagai *library* utama dalam membangun model XGBoost Regression. *Library* ini mendukung algoritma *gradient boosting* yang dirancang untuk menghasilkan model prediktif yang kuat pada data tabular. XGBoost digunakan untuk membangun model pembanding terhadap Random Forest Regression. Penggunaan *library* XGBoost diperlukan karena model yang diuji dalam penelitian secara khusus mencakup algoritma XGBoost Regression.

2.4.3 Matplotlib dan Seaborn

Matplotlib dan Seaborn adalah *library* visualisasi data dalam penelitian ini. Visualisasi diperlukan untuk mendukung tahap *data understanding, preprocessing,*

dan evaluation. Melalui visualisasi, peneliti dapat memahami distribusi data, melihat pola hubungan antarvariabel, mengidentifikasi nilai ekstrem, serta menyajikan hasil evaluasi model secara lebih mudah.

Matplotlib digunakan karena memiliki kemudahan dalam membuat berbagai bentuk grafik, seperti histogram, bar chart, line plot, dan visualisasi perbandingan hasil model. Matplotlib juga dapat digunakan untuk menampilkan distribusi nilai pasar pemain, perbandingan hasil evaluasi model, serta visualisasi hasil prediksi. Sementara itu, Seaborn digunakan untuk mendukung visualisasi statistik yang lebih ringkas dan mudah dibaca, seperti heatmap korelasi, boxplot, dan distribusi data.

2.4.4 Microsoft Word dan Reference Manager

Microsoft Word digunakan sebagai perangkat utama dalam penyusunan laporan skripsi. *Software* ini digunakan untuk menulis naskah penelitian, menyusun struktur bab dan subbab, mengatur heading, membuat tabel, memasukkan gambar, mengatur *caption*, serta menjaga konsistensi format dokumen sesuai dengan ketentuan akademik. Penggunaan Microsoft Word penting karena laporan skripsi harus disusun secara formal, rapi, dan mengikuti pedoman penulisan yang berlaku.

Selain Microsoft Word, penelitian ini juga menggunakan *reference manager*, seperti Mendeley atau Zotero, untuk membantu pengelolaan sitasi dan daftar pustaka. Reference manager digunakan untuk menyimpan referensi, mengatur metadata sumber, menyisipkan sitasi ke dalam dokumen, serta menyusun daftar pustaka secara lebih konsisten. Penggunaan *reference manager* penting karena penelitian ini menggunakan berbagai sumber akademik, seperti jurnal ilmiah, buku, artikel, dan dokumentasi resmi yang perlu dikelola secara sistematis.

Secara keseluruhan, tools dan software yang digunakan dalam penelitian ini memiliki peran yang saling melengkapi. Python dan Jupyter Notebook digunakan sebagai lingkungan utama analisis. Pandas dan NumPy digunakan untuk pengolahan data dan komputasi numerik. Scikit-learn dan XGBoost digunakan untuk pembangunan serta evaluasi model machine learning. Matplotlib dan Seaborn digunakan untuk visualisasi data dan hasil penelitian. Sementara itu, Microsoft Word dan reference manager digunakan untuk penyusunan laporan serta pengelolaan referensi akademik.