

BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Tabel 2.1 Tabel Penelitian Terdahulu

No.	Nama Jurnal	Penulis	Tahun	Skor RMSE dan R ²	Hasil
1.	XGBoost vs. Linear Regression in Healthcare Call Volume Prediction	M. M. T. Ayyalasomayajula, et al.	2023	Yearly: MSE LR=17.14, XGB=15.60; MAE LR=2.86, XGB=2.91	Regresi linear sedikit lebih baik pada MAE; XGBoost lebih baik menangkap puncak dan periode volatil seperti Natal.
2.	Predicting Football Transfer Fees Using Machine Learning	J. Nilsson, M. Andersson	2022	RMSE = €2.89 million; R ² = 0.88	Model penumpukan mengungguli XGBoost saja; semua model lebih baik daripada estimasi nilai pasar untuk memprediksi biaya transfer sepak bola.
3.	Predicting NHL Salaries with Ridge, LASSO, and XGBoost Regression	B. Klassen	2022	RMSE (LASSO) = 1,447,595	Model regresi LASSO memiliki RMSE terendah, yang menunjukkan akurasi prediksi gaji yang lebih baik daripada XGBoost dan model nonlinier lainnya.
4.	Perbandingan XGBoost dan Logistic Regression	S. E. H. Yulianti, O. Soesanto, Y. Sukmawaty	2025	Akurasi: XGBoost 97% (CI: 96,25%-97,73%),	XGBoost lebih unggul dalam menangkap pola non-linear

	dalam Memprediksi Credit Card Customer Churn			Logistic Regression 85% (CI: 83,17%-86,23%)	dan mengurangi overfitting, menghasilkan akurasi dan stabilitas lebih tinggi.
5.	Salary Prediction in Sports Analytics: MLB and NHL Cases	D. Drake, W. Smith	2024	MLB XGBoost $R^2 \approx 0.70$; NHL LASSO RMSE ≈ 1.45 M USD	Regresi XGBoost dan LASSO memberikan hasil prediksi gaji terbaik dalam kumpulan data bisbol dan hoki.
6.	Housing Price Prediction: XGBoost vs. Linear Regression	E. S. Putra, F. R. Nugroho	2023	RMSE LR=401,279; XGB=278,713; R^2 LR=0.723, XGB=0.862	XGBoost mengungguli regresi linier dan hutan acak dengan R^2 yang lebih tinggi dan RMSE yang lebih rendah dalam prediksi harga perumahan.
7.	Laptop Price Prediction Using XGBoost and Ensemble Methods	S. R. Putra, A. H. Nugroho	2022	$R^2=0.9277$; RMSE=221.66	XGBoost mencapai R^2 tertinggi dan RMSE terendah dibandingkan model lain dalam prediksi harga laptop.
8.	Energy Consumption Forecasting Using XGBoost and Linear Regression	L. Wang, M. Chen	2022	RMSE = 0.15 (XGBoost), R^2 = 0.92; RMSE = 0.22, R^2 = 0.85	XGBoost memberikan prediksi yang lebih akurat dengan R^2 yang lebih tinggi, sedangkan regresi linear kurang mampu menangkap fluktuasi musiman.
9.	Crop Yield Prediction Using Machine	A. Gupta, P. Singh	2024	RMSE = 3.8 ton/ha (XGBoost), R^2 =	XGBoost menunjukkan performa lebih

	Learning Techniques			0.84; RMSE = 4.5 ton/ha, $R^2 = 0.72$	baik dalam menangkap variabilitas hasil panen dibandingkan regresi linear.
10.	Predicting Student Performance Using Machine Learning Algorithms	S. K. Sharma, R. Kumar	2023	RMSE = 4.12, $R^2 = 0.78$ (XGBoost); RMSE = 5.05, $R^2 = 0.65$	XGBoost menunjukkan performa lebih baik dibanding regresi linear dengan RMSE lebih rendah dan R^2 lebih tinggi, menandakan kemampuan model menangkap pola non-linear dalam data pendidikan.

Berbagai penelitian terdahulu telah mengangkat topik prediksi gaji maupun prediksi variabel numerik lainnya dengan menggunakan algoritma machine learning seperti XGBoost dan regresi linear. Misalnya, Klassen (2017) dalam prediksi gaji pemain NHL menemukan bahwa regresi LASSO memberikan nilai RMSE terendah sekitar 1,45 juta USD, mengungguli XGBoost dan model lainnya. Sementara itu, Nilsson dan Andersson (2022) menggunakan XGBoost dan model ensemble untuk memprediksi biaya transfer pemain sepak bola, dengan hasil R^2 sebesar 0,88 dan RMSE sekitar €2,89 juta, menunjukkan keunggulan pendekatan ensemble dalam konteks pasar sepak bola. Dalam konteks non-olahraga, Putra dan Nugroho (2022) memprediksi harga laptop dan melaporkan bahwa XGBoost mencapai R^2 tertinggi sebesar 0,9277 dengan RMSE terendah 221,66, lebih baik dibandingkan Random Forest dan Gradient Boosting.

Ayyalasomayajula et al. (2023) membandingkan XGBoost dan regresi linear dalam memprediksi volume panggilan layanan kesehatan dan menemukan bahwa regresi linear lebih unggul dalam MAE, namun XGBoost lebih baik dalam menangkap pola fluktuatif. Zhang (2024) mengembangkan model regresi linear berganda untuk memprediksi gaji pemain NBA berdasarkan statistik performa, sedangkan Satpute et al. (2025) menunjukkan bahwa XGBoost unggul dalam klasifikasi gaji dibandingkan regresi logistik dan random forest.

Berbeda dari penelitian-penelitian tersebut yang umumnya hanya menggunakan satu model utama atau fokus pada sektor non-baseball, penelitian ini membandingkan secara langsung dua pendekatan populer—Linear Regression (sebagai baseline) dan XGBoost (sebagai model non-linear yang kompleks)—dalam konteks prediksi gaji pemain baseball berdasarkan statistik performa. Perbedaan utama solusi yang ditawarkan dalam penelitian ini adalah fokus pada integrasi kedua model dalam satu studi untuk mengevaluasi keunggulan masing-masing secara empiris pada domain baseball, serta penerapan preprocessing khusus termasuk penyesuaian inflasi dan pemilihan fitur berbasis performa aktual. Dengan demikian, penelitian ini tidak hanya memperluas penerapan algoritma pada domain olahraga yang berbeda, tetapi juga menawarkan pendekatan evaluatif yang lebih komprehensif terhadap akurasi dan relevansi model dalam konteks gaji berbasis performa.

2.2 Teori Penelitian

2.2.1 Machine Learning dalam Prediksi Olahraga

Analisis data dalam olahraga, terutama baseball, telah menjadi elemen penting dalam memahami dan memprediksi kinerja pemain. Baseball menghasilkan berbagai data statistik yang kaya dan terperinci, mulai dari hitungan strike, home run, hingga rata-rata pukulan dan persentase keberhasilan pitcher. Dengan memanfaatkan data ini, pelatih, manajer tim, dan analis dapat menggali wawasan mendalam mengenai pola permainan dan faktor-faktor kunci yang memengaruhi performa pemain. Dalam konteks prediksi, analisis data memungkinkan identifikasi tren dan pola yang sulit diobservasi secara manual. Misalnya, model prediksi dapat digunakan untuk memperkirakan performa pemain di masa depan atau untuk memilih strategi terbaik berdasarkan kondisi spesifik pertandingan. Penggunaan algoritma seperti XGBoost dan regresi linear dapat membantu mengolah data yang kompleks, mengoptimalkan akurasi prediksi, dan membantu tim membuat keputusan yang lebih efektif dan berbasis data. Analisis ini tidak hanya menguntungkan tim dalam mengatur strategi tetapi juga meningkatkan pengalaman penonton dengan menyediakan informasi mendalam tentang atlet dan permainan.

2.2.2 Baseball

Baseball merupakan olahraga beregu yang dimainkan oleh dua tim, masing-masing beranggotakan sembilan pemain. Permainan ini berlangsung selama sembilan inning, di mana setiap tim akan bergantian menjadi tim penyerang (offense) dan bertahan (defense). Tim penyerang berusaha mencetak run dengan memukul bola yang dilempar oleh pitcher, lalu berlari mengelilingi keempat base (first, second, third, dan home plate) hingga kembali ke home plate [28].

Sejarah baseball modern dimulai dari permainan rakyat Eropa seperti rounders dan cricket, kemudian dibakukan melalui aturan yang dibuat oleh Alexander Cartwright pada pertengahan abad ke-19. MLB profesional pertama, Cincinnati Red Stockings, dibentuk pada tahun 1869 [29]. Sejak itu, baseball telah menyebar ke Jepang, Korea, Amerika Latin, dan menjadi olahraga global dengan turnamen seperti MLB, World Baseball Classic, dan kadang hadir di Olimpiade.

Selain aspek teknis seperti melempar (pitching), memukul (hitting), dan menangkap (fielding), baseball juga menampilkan dimensi statistik yang sangat mendalam. Lewat dukungan teknologi seperti Statcast, permainan kini dianalisis dari sudut biomekanik, kinematik, serta pengaruh kondisi lingkungan seperti suhu terhadap jarak home run.

Salah satu faktor penting yang memengaruhi besaran gaji pemain baseball adalah posisi bermain. Setiap posisi memiliki peran strategis dan tingkat kesulitan yang berbeda, sehingga berdampak pada nilai kontrak yang ditawarkan. Misalnya, posisi seperti pitcher dan catcher cenderung memperoleh gaji lebih tinggi karena peran mereka sangat krusial dalam pertahanan dan pengendalian permainan. Sementara itu, pemain dengan posisi middle infielder seperti shortstop atau second baseman sering dinilai berdasarkan kombinasi performa ofensif dan defensif, yang turut memengaruhi nilai ekonomi mereka. Studi oleh Meltzer dan Fort (2018) menunjukkan bahwa posisi memiliki kontribusi signifikan terhadap struktur gaji di Major League Baseball (MLB),

dengan hasil bahwa posisi premium seperti pitcher dan first basemen memperoleh rata-rata gaji lebih tinggi dibandingkan posisi lain [46].

Selain faktor peran, kelangkaan pemain berkualitas di suatu posisi juga turut meningkatkan nilai pasar. Penelitian oleh Gassmann dan Bruggink (2020) menunjukkan bahwa ketika pasokan pemain berbakat di suatu posisi terbatas, tim-tim cenderung bersedia membayar lebih tinggi demi mengamankan talenta tersebut [47]. Oleh karena itu, dalam proses prediksi gaji, informasi posisi pemain menjadi fitur penting yang harus dianalisis secara kuantitatif. Dengan menerapkan algoritma pembelajaran mesin seperti XGBoost, perbedaan nilai kontribusi antar posisi dapat diukur melalui pentingnya fitur (feature importance), yang memperkuat pemahaman akan bagaimana posisi memengaruhi evaluasi dan kompensasi pemain secara keseluruhan.

2.2.3 Major League Baseball

Major League Baseball (MLB) adalah liga bisbol profesional tertinggi di Amerika Serikat dan Kanada yang terdiri dari 30 tim, terbagi dalam dua liga utama: American League (AL) dan National League (NL). Didirikan sejak akhir abad ke-19, MLB telah berkembang menjadi salah satu liga olahraga paling bergengsi dan berpengaruh di dunia. Setiap musimnya, MLB menyelenggarakan ratusan pertandingan yang memuncak dalam ajang bergengsi World Series, di mana dua tim terbaik dari masing-masing liga bertanding untuk memperebutkan gelar juara.

MLB tidak hanya menjadi ajang kompetisi olahraga, tetapi juga merupakan industri besar yang mencakup hak siar, sponsor, penjualan merchandise, serta pengelolaan stadion dan basis penggemar. Dalam beberapa dekade terakhir, liga ini

juga semakin memanfaatkan teknologi, termasuk analitik data, sistem pelacakan performa, dan tayangan ulang berbasis video untuk mendukung berbagai aspek permainan dan manajemen tim. Selain itu, MLB memainkan peran penting dalam budaya populer Amerika dan menjadi wadah bagi para atlet dari berbagai negara untuk menunjukkan kemampuan mereka di panggung internasional.

2.2.4 Prediksi Gaji

Dalam melakukan analisis prediksi gaji pemain baseball dari waktu ke waktu, perlu memperhitungkan inflasi agar nilai gaji yang digunakan bersifat realistis dan sebanding. Untuk tujuan ini, penyesuaian gaji dilakukan setiap Januari menggunakan indeks harga konsumen (Consumer Price Index, CPI). Rumus yang digunakan adalah:

$$P_t^{Real} = P_t^{Nominal} \times \frac{CPI\ January\ Base}{CPI\ January\ t}$$

Keterangan :

P_t^{Real} : Gaji riil (real salary) pemain pada tahun ke- t setelah disesuaikan terhadap inflasi.

$P_t^{Nominal}$: Gaji nominal pemain pada tahun ke- t tanpa penyesuaian inflasi.

$CPI\ January\ Base$: Nilai indeks harga konsumen (CPI) pada bulan Januari tahun dasar yang dijadikan acuan (misalnya tahun 2024 atau 2025).

$CPI\ January\ t$: Nilai indeks harga konsumen (CPI) pada bulan Januari tahun ke- t atau tahun data gaji tersebut tercatat.

Penerapan penyesuaian gaji ini menjamin bahwa model prediksi menangkap nilai ekonomi aktual dan meminimalkan bias akibat perubahan harga secara umum. Dengan data gaji yang telah disesuaikan, selanjutnya metode machine learning seperti XGBoost dan Linear Regression dapat digunakan untuk memprediksi gaji riil berdasarkan variabel performa pemain tanpa terganggu efek inflasi.

2.3 Framework dan Algoritma Penelitian

2.3.1 Cross-Industry Standard Process for Data Mining (CRISP-DM)

Dalam penelitian ini, proses pengolahan dan analisis data mengikuti pendekatan CRISP-DM (Cross-Industry Standard Process for Data Mining), yang merupakan salah satu kerangka kerja yang paling banyak digunakan dalam proyek data science dan machine learning. CRISP-DM dikembangkan pada akhir 1990-an oleh konsorsium yang terdiri dari Daimler-Benz, SPSS, dan NCR, dan hingga kini tetap relevan karena pendekatannya yang fleksibel dan dapat diterapkan di berbagai industri serta topik penelitian [42].

CRISP-DM terdiri dari enam tahapan utama yang saling berkaitan dan iteratif, yaitu:

1. Business Understanding
Tahap pertama ini bertujuan untuk memahami tujuan bisnis atau latar belakang permasalahan yang ingin diselesaikan. Dalam konteks penelitian ini, tujuan bisnis diterjemahkan ke dalam bentuk akademik, yakni membangun model prediksi gaji pemain baseball berdasarkan statistik performa yang tersedia. Pemahaman terhadap

domain baseball serta struktur kontrak pemain diperlukan untuk merumuskan pertanyaan penelitian yang tepat.

2. Data Understanding

Setelah tujuan dirumuskan, langkah berikutnya adalah memahami data yang tersedia. Ini meliputi pengumpulan data dari berbagai sumber seperti Fangraphs, Baseball-Reference, dan Sean Lahman Database, serta data ekonomi dari Bureau of Labor Statistics (BLS). Proses ini mencakup eksplorasi awal terhadap struktur, tipe data, distribusi nilai, dan deteksi masalah potensial seperti missing value atau outlier.

3. Data Preparation

Tahap ini merupakan proses paling intensif dalam data science, karena mencakup pembersihan data, transformasi, penggabungan beberapa dataset, encoding variabel kategorikal, normalisasi fitur numerik, dan seleksi fitur yang relevan. Dalam penelitian ini, juga dilakukan penyesuaian nilai gaji menggunakan indeks inflasi tahunan (CPI) agar gaji pada tahun yang berbeda dapat dibandingkan dalam satuan nilai riil. Data preparation memiliki pengaruh besar terhadap performa model yang akan dibangun.

4. Modeling

Setelah data siap, proses pemodelan dilakukan menggunakan dua algoritma: Linear Regression sebagai baseline model karena kesederhanaannya dan interpretabilitasnya yang tinggi, serta XGBoost sebagai model berbasis ensemble learning yang lebih

kompleks dan mampu menangani data dengan hubungan non-linear. Pemilihan parameter model dan proses tuning hyperparameter juga dilakukan di tahap ini untuk meningkatkan akurasi prediksi.

5. Evaluation

Tahap ini bertujuan untuk mengevaluasi kinerja model yang telah dibangun. Pengujian dilakukan dengan membandingkan hasil prediksi terhadap data aktual menggunakan metrik evaluasi seperti Root Mean Square Error (RMSE), Mean Absolute Error (MAE), dan koefisien determinasi (R^2). Evaluasi ini menentukan apakah model telah memenuhi tujuan penelitian dan memberikan hasil yang dapat diandalkan.

6. Deployment

Dalam konteks bisnis, tahap deployment biasanya mencakup implementasi model ke sistem produksi. Namun, dalam konteks penelitian akademik, deployment diterjemahkan sebagai penyajian hasil penelitian, interpretasi model, serta penarikan kesimpulan dan rekomendasi berdasarkan hasil analisis. Hasil akhir dapat berupa visualisasi prediksi, fitur paling berpengaruh terhadap gaji, dan usulan penggunaan model dalam pengambilan keputusan manajemen pemain.

Kerangka CRISP-DM tidak hanya menyediakan urutan tahapan yang sistematis, tetapi juga menekankan pentingnya feedback loop—artinya, setiap tahap dapat kembali ke tahap sebelumnya jika diperlukan. Hal ini sangat berguna dalam penelitian berbasis data yang bersifat eksploratif dan dinamis.

Dengan menggunakan CRISP-DM sebagai acuan, penelitian ini dapat dilakukan secara lebih terstruktur dan berorientasi pada tujuan akhir, yakni menghasilkan model prediksi gaji pemain baseball yang akurat dan aplikatif. Fleksibilitas dan kestabilan metodologi ini juga menjadikan CRISP-DM sebagai pilihan yang sesuai untuk penelitian yang menggabungkan data numerik, domain olahraga, dan teknik pembelajaran mesin.

2.3.2 XGBoost

XGBoost, atau Extreme Gradient Boosting, adalah algoritma machine learning berbasis ensemble yang mengimplementasikan teknik boosting untuk meningkatkan akurasi prediksi. Algoritma ini bekerja dengan membangun model secara bertahap, di mana setiap model baru berusaha memperbaiki kesalahan dari model sebelumnya. XGBoost menggunakan pendekatan gradient boosting, yang berarti ia mengoptimalkan kesalahan menggunakan teknik gradien untuk meminimalkan selisih antara hasil prediksi dan nilai aktual. Keunggulan utama XGBoost adalah kemampuannya untuk menangani data yang besar, kompleks, dan berdimensi tinggi, serta kemampuannya untuk mengurangi overfitting melalui penggunaan regularisasi yang ketat[10].

$$\hat{y}_i = \sum_{t=1}^T F_t(x_i)$$

Di mana:

- $\hat{y}_i$ = Prediksi akhir untuk data ke- i .
- T = Jumlah total pohon keputusan.

- $F_t(x_i)$ = Prediksi dari pohon ke-t untuk input x_i .

Tree Boosting Tree boosting adalah inti dari XGBoost, di mana algoritma ini menggabungkan banyak pohon keputusan (decision trees) yang dibangun secara berurutan. Setiap pohon baru difokuskan untuk memperbaiki kesalahan dari pohon sebelumnya. XGBoost menggunakan pendekatan additive training, di mana model yang dihasilkan adalah akumulasi dari semua pohon yang telah dibangun. Proses ini menghasilkan model akhir yang lebih kuat dan akurat, karena setiap pohon berkontribusi terhadap perbaikan performa secara keseluruhan. Dalam konteks data yang rumit seperti baseball, XGBoost mampu mengidentifikasi interaksi non-linear antara variabel-variabel statistik pemain yang mungkin tidak dapat ditangkap oleh algoritma sederhana.

Regularisasi Salah satu fitur utama dari XGBoost adalah penerapan teknik regularisasi yang ketat, yang bertujuan untuk mencegah overfitting, yaitu situasi di mana model terlalu “menghafal” data latih dan gagal menghasilkan prediksi yang baik pada data baru. XGBoost menerapkan dua jenis regularisasi:

L1 Regularization (Lasso): Membantu dalam penyederhanaan model dengan menekan nilai beberapa koefisien menjadi nol, sehingga variabel-variabel yang kurang signifikan diabaikan[11].

L2 Regularization (Ridge): Membatasi ukuran parameter untuk menghindari koefisien yang terlalu besar, membantu membuat model lebih stabil dan generalisasi pada data baru[11].

Regularisasi ini sangat bermanfaat ketika bekerja dengan data yang kompleks, karena membantu menjaga keseimbangan antara akurasi dan kemampuan generalisasi model.

2.3.3 Linear Regression

Regresi linear adalah metode statistik yang digunakan untuk memodelkan hubungan antara variabel bebas (independen) dan variabel tergantung (dependen). Dalam regresi linear sederhana, model mencari garis lurus terbaik yang mendekati data dengan meminimalkan selisih antara nilai yang diprediksi dan nilai aktual[12]. Secara umum, regresi linear merepresentasikan hubungan ini dengan persamaan:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Dimana:

- y : Variabel dependen (variabel output atau respons).
- x : Variabel independen (variabel input atau prediktor).
- β_0 : Intersep (nilai y ketika $x = 0$).
- β_1 : Gradien atau kemiringan (tingkat perubahan y terhadap x).
- ϵ : Error (kesalahan yang mencakup variabilitas yang tidak dijelaskan oleh model).

Dalam pemodelan statistik, persamaan regresi yang telah diestimasi (tanpa error) sering ditulis sebagai:

$$\hat{y} = \beta_0 + \beta_1 x$$

Di mana $\hat{y}$ adalah nilai prediksi dari y .

Salah satu asumsi dasar dalam regresi linear adalah linearitas, di mana diasumsikan bahwa hubungan antara variabel bebas dan variabel terikat bersifat linear atau dapat direpresentasikan dengan garis lurus. Ketika diterapkan pada konteks performa pemain, linearitas berarti bahwa perubahan dalam variabel-variabel tertentu (seperti jumlah pukulan atau rata-rata pukulan dalam baseball) dapat diprediksi menghasilkan perubahan proporsional dalam kinerja keseluruhan. Misalnya, jika rata-rata pukulan seorang pemain meningkat, performa mereka secara keseluruhan mungkin juga meningkat dengan pola yang cukup linear.

Namun, untuk hubungan yang sangat kompleks, regresi linear mungkin memiliki keterbatasan, dan model non-linear atau teknik boosting seperti XGBoost mungkin lebih sesuai. Meski demikian, regresi linear tetap sangat berguna ketika variabel menunjukkan hubungan linear yang jelas atau ketika hasil yang dapat diinterpretasikan dengan mudah lebih diutamakan.

Dalam skripsi ini, penggunaan model XGBoost dan Linear Regression dipilih untuk membandingkan kemampuan dalam menangani hubungan linier dan non-linier pada prediksi gaji pemain. XGBoost diterapkan karena kemampuannya menangkap pola kompleks dan interaksi antar variabel melalui teknik boosting dan regularisasi yang kuat, menghasilkan nilai R^2 dan RMSE yang unggul dibanding metode linier klasik [40]. XGBoost juga menawarkan stabilitas terhadap outlier dan variabel yang hilang, serta otomatisasi dalam pemilihan fitur penting—menjadikannya sangat sesuai untuk dataset statistik pemain yang kaya. Sementara itu, Linear Regression digunakan sebagai model referensi sederhana dan mudah

diinterpretasi; koefisiennya langsung menunjukkan pengaruh tiap fitur terhadap gaji, yang mendukung pendekatan analisis inferensial yang sering dibutuhkan dalam studi akademik[41]. Dengan membandingkan kedua model ini, penelitian ini tidak hanya mengevaluasi akurasi prediksi tetapi juga mengeksplorasi trade-off antara interpretabilitas dan performa prediksi.

2.3.4 RMSE

Root Mean Square Error (RMSE) adalah metrik evaluasi yang digunakan untuk mengukur seberapa besar kesalahan antara nilai yang diprediksi oleh model dengan nilai aktual. RMSE memberikan informasi seberapa jauh prediksi model dari nilai sebenarnya dalam satuan yang sama dengan variabel target[13].

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Keterangan:

- y_i : Nilai aktual untuk observasi ke- i
- $\hat{y}_i$: Nilai prediksi untuk observasi ke- i
- n : Jumlah total observasi

Semakin kecil nilai RMSE, maka model semakin akurat. RMSE sangat sensitif terhadap outlier karena menggunakan kuadrat dari selisih.

2.3.5 R²

R-squared (R²) adalah ukuran statistik yang menunjukkan seberapa besar proporsi variansi dalam variabel dependen yang dapat dijelaskan oleh variabel independen dalam model. Nilai R² berada antara 0 dan 1[14].

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$$

Keterangan:

- y_i : Nilai aktual
- $\hat{y}_i$: Nilai prediksi
- $\bar{y}_i$: Rata-rata dari nilai aktual
- Pembilang: Total squared error
- Penyebut: Total variance dari data aktual

Semakin tinggi nilai R² (mendekati 1), semakin baik model dalam menjelaskan variabilitas data. Nilai R² = 0 berarti model tidak menjelaskan variansi apa pun.

2.3.6 Feature Selection

Untuk meningkatkan performa dan efisiensi model prediksi gaji pemain bisbol, dilakukan tahap feature selection untuk menyaring fitur-fitur paling relevan dari total 121 atribut statistik. Metode yang digunakan adalah SelectKBest dari scikit-learn, yang memilih fitur berdasarkan skor uji statistik (seperti ANOVA F-value) terhadap variabel target (gaji). Proses ini dilakukan secara iteratif dengan

mencoba berbagai nilai k , serta divalidasi menggunakan cross-validation. Hasilnya, dipilih 9 fitur terbaik yang secara signifikan membantu menjelaskan variansi target, sesuai prinsip “rule of thumb” (~5–10% dari total fitur) untuk menjaga keseimbangan antara kompleksitas dan interpretabilitas model—sejalan dengan panduan praktis di bidang feature selection [43], [44]. Seleksi ini juga efektif dalam mengurangi masalah curse of dimensionality, meningkatkan kecepatan pelatihan, dan menjaga generalisasi model terhadap data baru [45]. Dengan 9 fitur tersaring, model menjadi lebih sederhana, cepat, dan terhindar dari overfitting, sambil tetap mempertahankan akurasi prediksi yang tinggi.

2.4 Teori tentang tools/software yang digunakan

2.4.1 Python

Python merupakan bahasa pemrograman yang dikenal karena kesederhanaannya, fleksibilitas, serta dukungan yang luas untuk berbagai aplikasi, mulai dari pengembangan web hingga analisis data dan kecerdasan buatan [14]. Bahasa ini dirancang dengan filosofi yang menekankan keterbacaan kode, sehingga memudahkan pembelajaran dan penggunaan baik oleh pemula maupun profesional. Salah satu keunggulan Python adalah sintaksnya yang sederhana dan mudah dipahami, menyerupai bahasa Inggris, yang mempercepat proses pengembangan aplikasi. Contohnya, perintah untuk menampilkan teks hanya membutuhkan satu baris kode, seperti `print("Hello, world!")`, yang langsung mencetak output di layar.

Selain kemudahan tersebut, Python memiliki ekosistem pustaka (library) dan framework yang sangat lengkap, memungkinkan pengembang menangani berbagai jenis aplikasi dengan lebih efisien. Di bidang ilmu data dan analisis,

library seperti Pandas, NumPy, dan SciPy banyak digunakan untuk manipulasi data dan analisis numerik. Python juga menjadi bahasa pilihan utama dalam kecerdasan buatan dan machine learning dengan dukungan library seperti Scikit-learn dan TensorFlow [15]. Dengan dukungan ekosistem ini, pengembang dapat fokus pada solusi masalah utama tanpa harus membangun semuanya dari awal.

2.4.2 Jupyter Notebook

Jupyter Notebook adalah sebuah platform pengembangan berbasis web yang interaktif, memungkinkan pengguna untuk menulis dan menjalankan kode sekaligus melihat hasilnya dalam satu interface yang terpadu. Meskipun mendukung berbagai bahasa pemrograman, Python adalah bahasa yang paling sering digunakan di dalamnya. Platform ini sangat ideal untuk kebutuhan analisis data, machine learning, dan visualisasi, karena memungkinkan eksekusi kode secara bertahap dengan hasil yang langsung terlihat. Selain itu, Jupyter Notebook memfasilitasi penggabungan kode dan teks dokumentasi dalam satu file, sehingga mempermudah pencatatan proses analisis dan hasil eksperimen yang dapat dengan mudah dibagikan atau dipublikasikan [17].

2.4.3 PyBaseball

Library pybaseball merupakan salah satu pustaka open-source dalam bahasa Python yang secara khusus dikembangkan untuk mendukung pengambilan dan analisis data bisbol dari berbagai sumber terpercaya seperti Baseball Reference, FanGraphs, dan Statcast. Library ini memungkinkan akses langsung ke data statistik pemain MLB, baik dari sisi batting, pitching, maupun data granular seperti kecepatan lemparan, launch angle, dan exit velocity. Dengan menggunakan fungsi

seperti `statcast()`, `playerid_lookup()`, dan `batting_stats()`, pengguna dapat mengakses data berkualitas tinggi untuk keperluan analisis performa pemain, prediksi kontrak, atau studi berbasis *sabermetrics*. Dalam praktiknya, *pybaseball* sering digunakan sebagai fondasi data dalam penelitian akademik dan analisis *machine learning* yang melibatkan prediksi gaji pemain, deteksi outlier performa, hingga visualisasi tren statistik dari musim ke musim [38]. Kelebihan lain dari pustaka ini adalah dukungan komunitas yang aktif, dokumentasi yang lengkap di GitHub, dan kemampuannya untuk diintegrasikan dengan pustaka analisis lainnya seperti *Pandas* dan *Scikit-learn*. Meskipun belum banyak dibahas dalam jurnal formal, *pybaseball* telah digunakan secara luas dalam berbagai proyek ilmiah dan eksperimen *data science* berbasis olahraga, termasuk dalam studi terkini tentang deteksi aksi dalam video baseball menggunakan data *Statcast* [39].

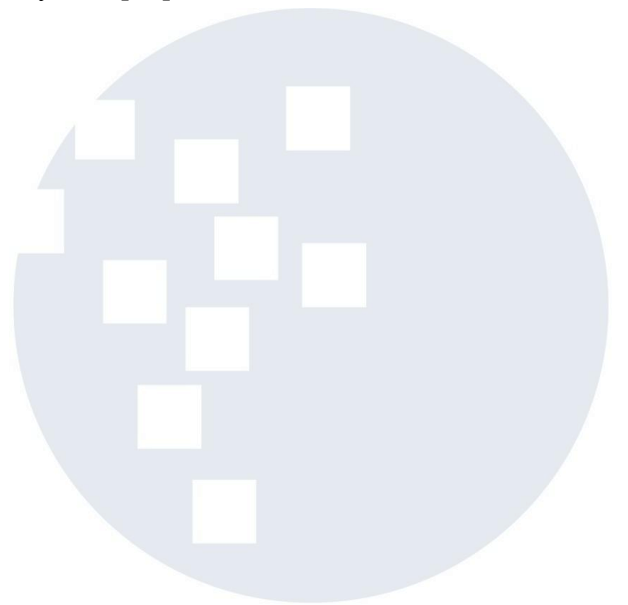
2.4.4 Streamlit

Streamlit adalah framework open-source berbasis Python yang digunakan untuk membangun aplikasi web interaktif dengan cepat dan sederhana, khususnya untuk keperluan analisis data dan *machine learning*. *Streamlit* memungkinkan pengguna untuk mengubah script Python menjadi antarmuka web tanpa perlu menulis kode HTML, CSS, atau JavaScript. Fitur-fitur seperti slider, dropdown, dan file uploader sangat membantu dalam membangun visualisasi data secara dinamis dan interaktif.

Dalam penelitian ini, *Streamlit* digunakan untuk menampilkan hasil analisis data dan prediksi model secara interaktif dalam bentuk dashboard. Hal ini

bertujuan agar hasil penelitian dapat diakses dan dipahami dengan lebih mudah oleh pengguna non-teknis.

Menurut dokumentasi resminya, Streamlit dirancang untuk para data scientist dan engineer agar dapat membuat prototipe aplikasi berbasis data hanya dengan beberapa baris kode Python [48].



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA