

BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

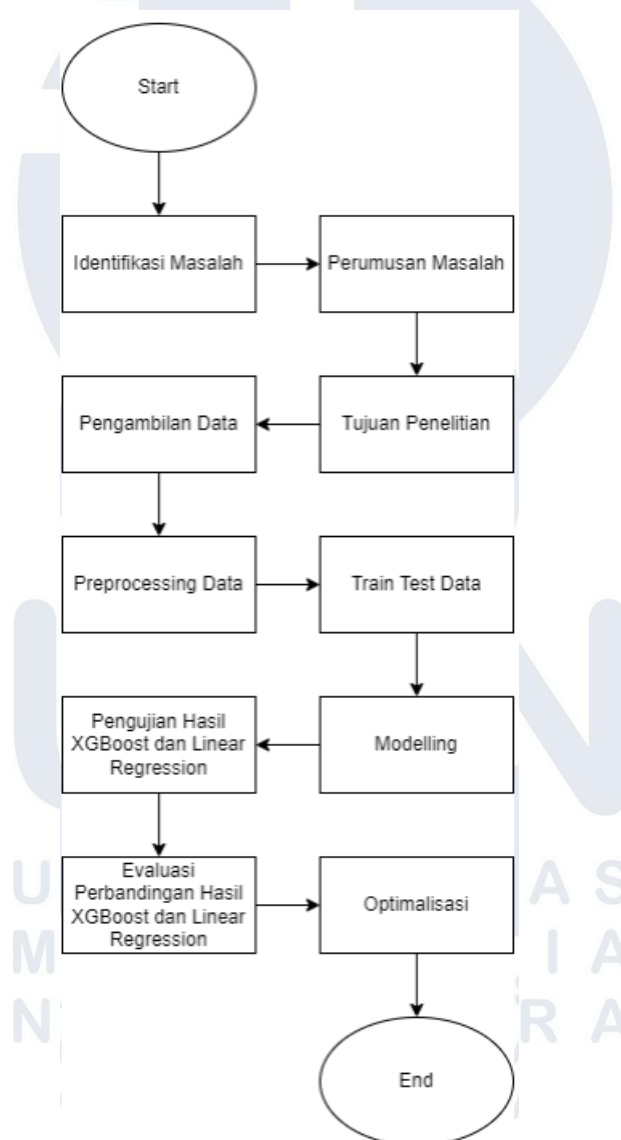
Objek penelitian dalam studi ini adalah olahraga baseball, khususnya aspek statistik pemain yang berkaitan dengan performa individu dan kontribusinya terhadap kesuksesan tim. Baseball dipilih karena memiliki struktur permainan yang terstandarisasi dan data statistik yang sangat lengkap seperti batting average, home run, dan earned run average. Data yang digunakan umumnya bersumber dari liga profesional seperti Major League Baseball (MLB), yang mendokumentasikan performa pemain secara sistematis dan historis dalam jumlah besar. Melalui pendekatan analisis data dan model prediktif, data statistik pemain ini dianalisis untuk mengidentifikasi variabel-variabel yang paling berpengaruh terhadap keberhasilan tim dalam suatu musim kompetisi. Dengan demikian, baseball bukan hanya menjadi objek kajian olahraga, tetapi juga sebagai representasi bidang yang kaya akan data untuk diterapkan dalam metode kuantitatif dan pembelajaran mesin.

3.2 Metode Penelitian

3.2.1 Alur Penelitian

Penelitian ini difokuskan pada pembangunan model prediktif untuk memperkirakan gaji pemain baseball berdasarkan data statistik performa pemain. Proses penelitian diawali dengan identifikasi masalah, yaitu belum optimalnya pendekatan yang digunakan dalam menentukan nilai kontrak atau gaji pemain yang berbasis pada analisis objektif performa. Setelah itu, dilakukan perumusan masalah yang bertujuan untuk menjawab

bagaimana model prediksi dapat dibangun dan algoritma mana yang memberikan hasil terbaik dalam memprediksi gaji pemain berdasarkan data performa historis. Tujuan penelitian ini adalah untuk membandingkan performa dua algoritma yaitu XGBoost dan Linear Regression dalam melakukan prediksi terhadap nilai gaji pemain, serta mengidentifikasi variabel statistik mana yang paling berpengaruh.



Gambar 3.1 Alur Penelitian

Data dikumpulkan dari sumber terbuka seperti Major League Baseball (MLB) dan situs data publik seperti Kaggle, yang memuat informasi statistik pemain seperti jumlah home run, batting average, RBI, WAR, dan metrik performa lainnya, serta nilai gaji tahunan masing-masing pemain. Setelah data diperoleh, dilakukan preprocessing yang meliputi pembersihan data, penanganan nilai kosong, encoding data kategorikal, normalisasi, dan seleksi fitur. Data yang telah siap kemudian dibagi menjadi data pelatihan dan data pengujian untuk menghindari overfitting dan memastikan validitas evaluasi model.

Tahap pemodelan dilakukan dengan menerapkan dua algoritma utama, yaitu Linear Regression dan XGBoost. Linear Regression dipilih sebagai baseline karena kesederhanaannya dan kemampuannya memberikan interpretasi langsung terhadap pengaruh setiap fitur terhadap target (gaji). Sementara itu, XGBoost digunakan sebagai model pembanding karena kemampuannya menangani kompleksitas hubungan antar fitur serta memiliki performa tinggi dalam prediksi berbasis data numerik. Proses tuning hyperparameter dilakukan untuk meningkatkan akurasi model, khususnya pada XGBoost.

Setelah model dibangun dan diuji, dilakukan evaluasi performa menggunakan metrik seperti Root Mean Square Error (RMSE), dan R-squared (R^2). Hasil dari kedua model kemudian dibandingkan untuk mengetahui algoritma mana yang lebih efektif dalam memprediksi gaji pemain berdasarkan statistik performa. Evaluasi ini juga menjadi dasar dalam memberikan rekomendasi terkait penggunaan pendekatan analitik dalam mendukung pengambilan keputusan kontraktual dalam industri baseball profesional.

3.3 Data Collection

Tahap Data Collection dilakukan untuk memahami karakteristik masing-masing dataset yang digunakan dalam penelitian ini. Sumber data yang digunakan berasal dari berbagai platform terpercaya yang

menyediakan statistik olahraga dan indikator ekonomi. Pertama, data Batting pemain individu diperoleh dari situs Fangraphs yang menyediakan informasi statistik pemukul pemain MLB sepanjang musim 2010 hingga 2024. Dataset ini terdiri dari 4.891 entri yang mencakup metrik seperti AVG (batting average), OBP (on-base percentage), SLG (slugging), HR (home run), dan WAR (wins above replacement). Data ini menjadi sumber utama dalam menentukan performa individu yang berkaitan langsung dengan nilai kontrak.

Kedua, data Fielding diperoleh dari Sean Lahman Baseball Database, yang merupakan repositori terbuka dan historis berisi statistik pertahanan (defensive metrics) pemain. Dataset ini mencakup 151.508 baris data yang mencatat kontribusi defensif pemain seperti posisi bermain, jumlah error, putouts, dan assists. Data ini digunakan untuk melengkapi gambaran performa total pemain, terutama bagi pemain non-pitcher.

Ketiga, data Team Batting dan Team Statistic diperoleh dari situs Baseball-Reference. Dataset Team Batting berisi 1.703 catatan performa kolektif tim dalam aspek ofensif, sedangkan Team Statistic mencakup 3.046 data terkait performa umum tim seperti jumlah kemenangan, peringkat, ERA tim, dan lainnya. Kedua dataset ini digunakan untuk memahami konteks tim tempat pemain berada, yang berpotensi memengaruhi nilai gaji atau kontrak mereka.

Keempat, data CPI (Consumer Price Index) tahunan Amerika Serikat dikumpulkan dari situs resmi U.S. Bureau of Labor Statistics. Data ini

mencakup 66 titik data CPI bulanan (khususnya Januari tiap tahun) dari periode 2010 hingga 2024, dan digunakan untuk menyesuaikan nilai kontrak atau gaji pemain terhadap inflasi, sehingga semua nilai gaji dapat dianalisis dalam satuan nilai riil.

Dalam tahap ini, dilakukan eksplorasi awal terhadap seluruh dataset untuk memahami struktur data, hubungan antar fitur, serta mengidentifikasi potensi permasalahan seperti missing values, outliers, ketidaksesuaian tipe data, dan inkonsistensi antar sumber. Pemahaman yang mendalam pada tahap ini menjadi fondasi penting untuk proses pembersihan dan pemodelan data pada tahap selanjutnya.

Selanjutnya proses joint data dilakukan untuk menggabungkan informasi statistik performa pemain dengan data gaji tahunan agar terbentuk dataset komprehensif yang dapat digunakan dalam pemodelan prediktif. Pada penelitian ini, digunakan teknik left join untuk menggabungkan data statistik pemain sebagai data utama (left table) dengan data gaji sebagai tabel referensi (right table) berdasarkan kunci penghubung berupa nama atau ID pemain dan tahun musim. Pemilihan metode left join bertujuan untuk mempertahankan seluruh data statistik pemain, termasuk pemain yang tidak memiliki informasi gaji. Hal ini penting dalam tahap eksplorasi awal dan analisis distribusi data, sekaligus membantu mengidentifikasi missing value yang mungkin memerlukan perlakuan lebih lanjut. Implementasi dilakukan menggunakan library pandas di Python dengan fungsi merge() dan parameter how='left'. Melalui pendekatan ini, data yang tersedia dapat diintegrasikan secara

menyeluruh tanpa menghilangkan informasi penting dari sisi statistik, sambil tetap menjaga fleksibilitas untuk penyesuaian data di tahap preprocessing berikutnya.

3.4 PreProcessing

1. Pemeriksaan Nilai Hilang (Missing Values)

Data pemain bisbol dikumpulkan dengan menggunakan library pybaseball dan dibagi menjadi statistik dasar dan lanjutan. Ada beberapa nilai gaji yang hilang, nilai-nilai ini diperhitungkan berdasarkan gaji sebelumnya yang tersedia untuk setiap pemain. Misalnya, jika Derek Jeter menghasilkan \$15 juta pada tahun 2010, dan ada nilai yang hilang untuk tahun 2011, nilai yang hilang akan diisi dengan gaji Derek Jeter tahun sebelumnya sebesar \$15 juta.

2. Pemeriksaan Duplikasi Data

Pemeriksaan duplikasi dilakukan untuk memastikan bahwa tidak terdapat data ganda yang dapat memengaruhi hasil analisis. Pemain yang berpindah tim pada saat musim berlangsung akan memiliki dua statistik yang sama

3. Penyesuaian Inflasi

Meskipun kumpulan data dibatasi dengan data dari 2010 hingga 2024, terdapat variasi dalam gaji pemain selama dua dekade terakhir. Untuk memperhitungkan hal ini, gaji pemain disesuaikan dengan inflasi menggunakan indeks CPI nasional Amerika.

4. Analisis Berdasarkan Posisi Pemain

Kumpulan data pemain bisbol dikelompokkan berdasarkan gaji rata-rata masing-masing pemain selama tahun-tahun pemain tersebut bermain. Metode ini secara efektif menghilangkan fitur kategoris seperti posisi dan tim yang dimainkan karena banyak pemain yang bermain di beberapa posisi dan tim selama karir mereka.

5. Analisis Korelasi

Untuk lebih memperhitungkan variabilitas dalam kontrak pemain antar tahun, perbedaan gaji pemain tiap tahun ditambahkan sebagai fitur ke dalam kerangka data pemain bisbol. Hal ini sebagian mengatasi perbedaan gaji yang besar bagi pemain yang menghasilkan lebih banyak saat memasuki masa free agency.

6. Feature Selection

Untuk mengurangi kompleksitas model, library `feature_selection` Sci-Kit Learn diimplementasikan dan menemukan bahwa metode `SelectKBest` memberikan kinerja terbaik saat mengidentifikasi fitur yang paling penting dan menjelaskan varian model.

7. Split Data

Setelah data selesai diproses, data dibagi menjadi dua bagian, yaitu:

- a. Data Training 80% dari total data, digunakan untuk melatih model.
- b. Data Testing 20% dari total data, digunakan untuk menguji performa model.
- c. 5-Fold Validation digunakan untuk mengevaluasi dataset.

Hasil dari tahap Data Preparation ini adalah dataset yang bersih sehingga siap digunakan dalam tahap pemodelan untuk membangun model prediksi berbasis XGBoost dan Linear Regression. Langkah ini memastikan bahwa model dapat belajar dari data yang relevan dan berkualitas sehingga meningkatkan performa model dan mengurangi error secara keseluruhan.

Dalam penelitian ini, dipilih metrik evaluasi Root Mean Squared Error (RMSE) dan R-squared (R^2). RMSE digunakan karena memberikan kesalahan prediksi dalam satuan yang sama dengan target (misalnya gaji), sehingga lebih mudah diinterpretasi secara praktis dibandingkan MSE, dan lebih sensitif terhadap outlier—ini penting dalam konteks gaji di mana kesalahan besar berdampak signifikan [33]. Sementara itu, R^2 mengukur proporsi variasi target yang berhasil dijelaskan oleh model, sehingga memberikan gambaran seberapa baik model menjelaskan hubungan antara fitur statistik dan gaji [34].

Pemilihan split data 80:20 (training:test) merupakan praktik mapan yang menyediakan keseimbangan antara data untuk pelatihan dan validasi; kombinasi ini telah terbukti menghasilkan estimasi performa yang stabil tanpa kehilangan kekuatan pelatihan secara signifikan [35]. Untuk meningkatkan keandalan, digunakan juga 5-fold cross-validation, yaitu teknik yang membagi data menjadi lima lipatan dan menguji model pada tiap lipatan secara bergantian. Literatur menunjukkan bahwa 5–10 fold merupakan kompromi optimal antara kompleksitas komputasi, bias, dan variansi estimasi error [36]. Selain itu, studi teoretis menyebutkan bahwa k-fold CV, termasuk nilai $k = 5$, menghasilkan

estimator out-of-sample yang konsisten ($\sqrt{n}$ -consistent) dan estimasi variansi yang valid [37].

3.5 Modelling

Pada tahap ini, akan dilakukan analisis terhadap pemodelan menggunakan algoritma XGBoost dan Linear Regression. Pemilihan algoritma ini berdasarkan penelitian terdahulu yang telah menggunakan XGBoost dan Linear Regression pada analisis statistik pemain bisbol maupun olahraga lainnya. Proses pemodelan akan menggunakan bahasa pemrograman Python untuk mengimplementasikan algoritma XGBoost dan Linear Regression secara efektif.

3.6 Evaluation

Pada langkah ini, dilakukan evaluasi kinerja dari model regresi yang telah dibangun. Evaluasi ini mencakup penyajian nilai R-squared (R^2) dan Root Mean Squared Error (RMSE) sebagai metrik evaluasi untuk mengukur seberapa baik model memprediksi nilai target. Model diuji menggunakan dua algoritma yaitu XGBoost Regression dan Linear Regression. Hasil evaluasi dari kedua algoritma tersebut kemudian dibandingkan berdasarkan nilai R^2 yang menunjukkan proporsi variansi yang dapat dijelaskan oleh model, serta nilai RMSE yang menggambarkan rata-rata besar kesalahan prediksi. Evaluasi ini bertujuan untuk menentukan model mana yang memberikan performa terbaik dalam melakukan prediksi terhadap data uji.

3.7 Deployment

Tahap terakhir dalam proses ini adalah menerapkan hasil yang didapatkan. Dalam penelitian ini, tahap implementasi akan mencakup pembuatan sebuah website yang akan menampilkan hasil prediksi nilai kontrak pemain bisbol terhadap data yang diinput. Website ini akan memberikan prediksi yang jelas terhadap data yang diperlukan oleh tim bisbol untuk melakukan akuisisi pemain.



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA