

**Pengembangan Model Prediksi Pemain Baseball Menggunakan  
Algoritma XGBoost dan Regresi Linier: Studi Kasus pada Data  
Statistik Pemain MLB**



**SKRIPSI**

**MUHAMMAD RIZKY AZZAKKY**

**00000033354**

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS TEKNIK DAN INFORMATIKA  
UNIVERSITAS MULTIMEDIA NUSANTARA  
TANGERANG  
2025**

**Pengembangan Model Prediksi Pemain Baseball Menggunakan  
Algoritma XGBoost dan Regresi Linier: Studi Kasus pada Data  
Statistik Pemain MLB**



**SKRIPSI**

**Diajukan sebagai Salah Satu Syarat untuk Memperoleh  
Gelar Sarjana Komputer**

**MUHAMMAD RIZKY AZZAKKY**

**00000033354**

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS TEKNIK DAN INFORMATIKA  
UNIVERSITAS MULTIMEDIA NUSANTARA  
TANGERANG**

**2025**

**i**

**ANALISA NILAI DAN STATISTIK PEMAIN BASEBALL UNTUK MEMPREDIKSI  
DAMPAK PEMAIN TERHADAP KESUKSESAN TIM, Muhammad Rizky Azzakky, Universitas  
Multimedia Nusantara**

## HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Muhammad Rizky Azzakky

Nomor Induk Mahasiswa : 00000033354

Program Studi : Sistem Informasi

Skripsi dengan judul:

**Pengembangan Model Prediksi Pemain Baseball Menggunakan Algoritma  
XGBoost dan Regresi Linier: Studi Kasus pada Data Statistik Pemain MLB**

Merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan **TIDAK LULUS** untuk mata kuliah yang telah saya tempuh.

Tangerang, 3 Juli 2025



(Muhammad Rizky Azzakky)

## HALAMAN PENGESAHAN

Skripsi dengan judul

**ANALISA NILAI DAN STATISTIK PEMAIN BASEBALL UNTUK  
MEMPREDIKSI DAMPAK PEMAIN TERHADAP KESUKSESAN TIM**

Oleh

Nama : Muhammad Rizky Azzakky  
NIM : 00000033354  
Program Studi : Sistem Informasi  
Fakultas : Teknik dan Informatika

Telah diujikan pada hari Kamis, 19 Juni 2025

Pukul 13.00 s.d 15.00 dan dinyatakan

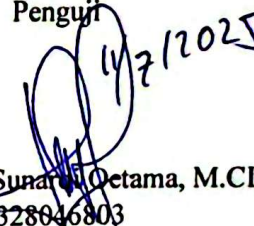
**LULUS**

Dengan susunan penguji sebagai berikut.

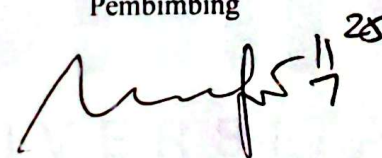
Ketua Sidang

  
Johan Setiawan. S.Kom., M.M.,  
MBA 327106402

Penguji

  
Ir. Raymond Sunardi Oetama, M.CIS  
0328046803

Pembimbing

  
Dr. Santo Fernandi Wijaya., S.Kom.,  
M.M. 081435

Ketua Program Studi Sistem Informasi

  
Ririn Ikana Desanti, S.Kom., M.Kom.

iv

**ANALISA NILAI DAN STATISTIK PEMAIN BASEBALL UNTUK MEMPREDIKSI  
DAMPAK PEMAIN TERHADAP KESUKSESAN TIM, Muhammad Rizky Azzakky, Universitas  
Multimedia Nusantara**

## HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

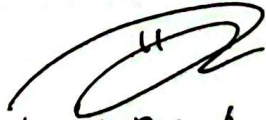
Yang bertanda tangan di bawah ini:

Nama : Muhammad Rizky Azzakky  
NIM : 00000033354  
Program Studi : Sistem Informasi  
Jenjang : S1  
Judul Karya Ilmiah : Analisa nilai dan statistik pemain baseball  
untuk memprediksi dampak pemain terhadap  
kesuksesan tim

Menyatakan dengan sesungguhnya bahwa saya bersedia\* (pilih salah satu):

- ☒ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- ☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) \*\*.
- ☐ Lainnya, pilih salah satu:
  - ☐ Hanya dapat diakses secara internal Universitas Multimedia Nusantara
  - ☐ Embargo publikasi karya ilmiah dalam kurun waktu 3 tahun.

Tangerang, 3 Juli 2025

  
(Muhammad Rizky Azzakky)

\* Pilih salah satu

\*\* Jika tidak bisa membuktikan LoA jurnal/HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk dipublikasikan ke KC UMN dan menjadi hak institusi UMN.

vi

ANALISA NILAI DAN STATISTIK PEMAIN BASEBALL UNTUK MEMPREDIKSI  
DAMPAK PEMAIN TERHADAP KESUKSESAN TIM, Muhammad Rizky Azzakky, Universitas  
Multimedia Nusantara



## KATA PENGANTAR

(Kata Pengantar dapat dikembangkan dan harus meliputi ucapan rasa syukur, tujuan pembuatan tugas akhir, ucapan terima kasih, dan harapan pada hasil Tugas Akhir ini.)

Mengucapkan terima kasih kepada:

1. Dr. Ir. Andrey Andoko, M.Sc, selaku Rektor Universitas Multimedia Nusantara.
2. Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Universitas Multimedia Nusantara.
3. Ririn Ikana Desanti, S.Kom., M.Kom, selaku Ketua Program Studi Universitas Multimedia Nusantara.
4. Dr. Santo Fernandi Wijaya., S.Kom., M.M., sebagai Pembimbing pertama yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.

semoga karya ilmiah ini dapat memberikan kontribusi yang positif bagi pengembangan ilmu pengetahuan, khususnya dalam bidang yang menjadi fokus penelitian ini.

Tangerang, 6 Juni 2025



Muhammad Rizky Azzakky

# **Pengembangan Model Prediksi Pemain Baseball Menggunakan Algoritma XGBoost dan Regresi Linier: Studi Kasus pada Data Statistik Pemain MLB**

## **ABSTRAK**

Penelitian ini bertujuan untuk melakukan eksplorasi data statistik pemain baseball guna mendukung pengembangan model prediksi nilai kontrak berdasarkan performa. Data diperoleh dari tiga sumber utama: Fangraphs, Baseball-Reference, dan Sean Lahman. Dataset mencakup lima kategori utama: 4.891 data Batting (2010–2024), 66 data Consumer Price Index (CPI), 151.508 data Fielding, 1.703 data Team Batting, dan 3.046 data Team Statistic. Tahap eksplorasi dilakukan untuk memahami struktur, kualitas, serta pola hubungan antar data.

Langkah eksplorasi meliputi pemeriksaan nilai hilang, duplikasi, distribusi nilai, dan outlier. Statistik deskriptif diterapkan untuk menilai performa pemain, sementara analisis korelasi dilakukan untuk mengidentifikasi keterkaitan antara metrik performa dan nilai kontrak (jika tersedia). Dataset juga dianalisis berdasarkan posisi pemain, serta disesuaikan dengan CPI untuk mempertimbangkan pengaruh inflasi terhadap nilai kontrak.

Penggabungan antar dataset dilakukan berdasarkan tahun dan playerID untuk memberikan pandangan yang menyeluruh terhadap kontribusi pemain. Hasil eksplorasi menunjukkan adanya variasi performa yang signifikan antar posisi serta potensi pengaruh faktor eksternal seperti inflasi terhadap valuasi pemain. Eksplorasi ini menjadi tahap awal yang penting dalam membangun model prediktif yang kuat, serta memberikan dasar strategis bagi manajemen tim dalam pengambilan keputusan berbasis data.

**Kata kunci:** *Baseball Prediction, Linear Regression, Machine Learning, Sabermetrics, Statistics, XGBoost.*

# **Pengembangan Model Prediksi Pemain Baseball Menggunakan Algoritma XGBoost dan Regresi Linier: Studi Kasus pada Data Statistik Pemain MLB**

## ***ABSTRACT (English)***

*This study aims to explore baseball player statistics to support the development of predictive models for contract value based on performance. Data were collected from three major sources: Fangraphs, Baseball-Reference, and Sean Lahman. The dataset includes five key categories: 4,891 Batting records (2010–2024), 66 Consumer Price Index (CPI) entries, 151,508 Fielding records, 1,703 Team Batting entries, and 3,046 Team Statistics.*

*The exploration process involved examining missing values, duplicates, value distributions, and outliers. Descriptive statistics were applied to assess player performance, and correlation analysis was conducted to explore relationships between performance metrics and contract value (if available). The data were also segmented by player position and adjusted using CPI to account for inflation's effect on contracts.*

*Datasets were merged based on player ID and year to provide a holistic view of individual contributions. The results revealed notable performance variations across positions and indicated the possible influence of external factors such as inflation on player valuation. This exploration forms a crucial foundation for building robust predictive models and offers strategic insights for team management in data-driven decision-making.*

**Keywords:** *Baseball Prediction, Linear Regression, Machine Learning, Sabermetrics, Statistics, XGBoost.*



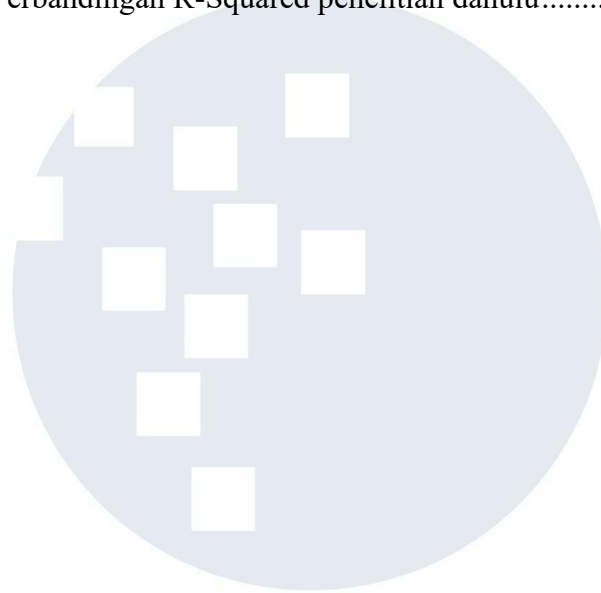
## DAFTAR ISI

|                                                                     |      |
|---------------------------------------------------------------------|------|
| HALAMAN PERSETUJUAN .....                                           | ii   |
| HALAMAN PENGESAHAN .....                                            | iii  |
| HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH .....                    | v    |
| KATA PENGANTAR .....                                                | vi   |
| ABSTRAK .....                                                       | vii  |
| <i>ABSTRACT (English)</i> .....                                     | viii |
| DAFTAR ISI .....                                                    | ix   |
| DAFTAR TABEL .....                                                  | xi   |
| DAFTAR GAMBAR .....                                                 | xii  |
| DAFTAR LAMPIRAN .....                                               | xiii |
| BAB I PENDAHULUAN .....                                             | 1    |
| 1.1 Latar Belakang .....                                            | 1    |
| 1.2 Rumusan Masalah .....                                           | 9    |
| 1.3 Batasan Masalah .....                                           | 9    |
| 1.4 Tujuan dan Manfaat Penelitian .....                             | 9    |
| 1.4.1 Tujuan Penelitian .....                                       | 9    |
| 1.4.2 Manfaat Penelitian .....                                      | 10   |
| 1.5 Sistematika Penulisan .....                                     | 10   |
| BAB II LANDASAN TEORI .....                                         | 14   |
| 2.1 Penelitian Terdahulu .....                                      | 14   |
| 2.2 Teori Penelitian .....                                          | 22   |
| 2.2.1 Machine Learning dalam Prediksi Olahraga .....                | 22   |
| 2.2.2 Baseball .....                                                | 23   |
| 2.2.3 Major League Baseball .....                                   | 25   |
| 2.2.4 Prediksi Gaji .....                                           | 25   |
| 2.3 Framework dan Algoritma Penelitian .....                        | 26   |
| 2.3.1 Cross-Industry Standard Process for Data Mining (CRISP-DM) .. | 26   |
| 2.3.2 XGBoost .....                                                 | 30   |
| 2.3.3 Linear Regression .....                                       | 32   |

|                                                       |           |
|-------------------------------------------------------|-----------|
| 2.3.4 RMSE .....                                      | 34        |
| 2.3.5 R2 .....                                        | 35        |
| 2.3.6 Feature Selection .....                         | 35        |
| 2.4 Teori tentang tools/software yang digunakan ..... | 36        |
| 2.4.1 Python .....                                    | 36        |
| <b>BAB III METODOLOGI PENELITIAN .....</b>            | <b>39</b> |
| 3.1 Gambaran Umum Objek Penelitian .....              | 39        |
| 3.2 Metode Penelitian .....                           | 39        |
| 3.3 Data Collection .....                             | 42        |
| 3.4 PreProcessing .....                               | 45        |
| 3.5 Modelling .....                                   | 47        |
| 3.6 Evaluation .....                                  | 48        |
| 3.7 Deployment .....                                  | 48        |
| <b>BAB IV ANALISIS DAN HASIL PENELITIAN .....</b>     | <b>50</b> |
| 4.1 PENGEMBANGAN MODEL .....                          | 50        |
| 4.2 DEPLOYMENT .....                                  | 79        |
| <b>BAB V SIMPULAN DAN SARAN .....</b>                 | <b>81</b> |
| 5.1 Simpulan .....                                    | 81        |
| 5.2 Saran .....                                       | 83        |
| <b>DAFTAR PUSTAKA .....</b>                           | <b>85</b> |
| <b>LAMPIRAN .....</b>                                 | <b>92</b> |

## DAFTAR TABEL

|                                                                                                      |    |
|------------------------------------------------------------------------------------------------------|----|
| Tabe 1.1 Perbandingan Nilai Organisasi pada 5 tim terkaya dan termiskin pada NFL, NBA, dan MLB ..... | 3  |
| Tabel 1.2 Perbandingan Pengeluaran pada 5 tim terkaya dan termiskin pada NFL, NBA, dan MLB .....     | 4  |
| Tabel 2.1 Tabel Penelitian Terdahulu .....                                                           | 14 |
| Tabel 4.1Tabel Perbandingan R-Squared penelitian dahulu.....                                         | 75 |



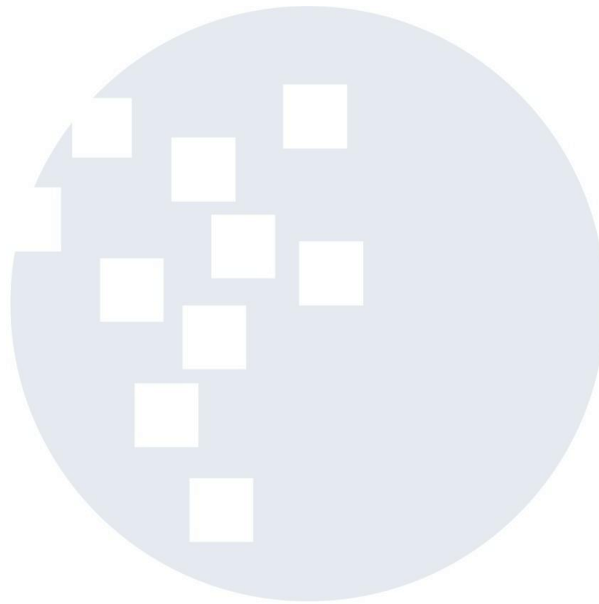
UMN  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

## DAFTAR GAMBAR

|                                                                           |    |
|---------------------------------------------------------------------------|----|
| Gambar 3.1 Alur Penelitian.....                                           | 41 |
| Gambar 4.2 Scrapping data batting_stats dari website Fangraph .....       | 50 |
| Gambar 4.3 Scrapping data bwar_bat dari website Baseball-References.....  | 51 |
| Gambar 4.4 Menggabungkan kedua data menggunakan left join.....            | 52 |
| Gambar 4.5 Scrapping data fielding dari website SeanLahman .....          | 52 |
| Gambar 4.6 Menggabungkan data fielding dengan data batting_merged.....    | 53 |
| Gambar 4.7 Mengisi nilai salary yang hilang .....                         | 54 |
| Gambar 4.8 Membuat Dataframe batting_basic .....                          | 55 |
| Gambar 4.9 Membuat Dataframe batting_adv.....                             | 56 |
| Gambar 4.10 Menghilangkan row dengan missng NaN .....                     | 56 |
| Gambar 4.11 Mencari data duplikat.....                                    | 56 |
| Gambar 4.12 Menghilangkan data duplikat .....                             | 57 |
| Gambar 4.13 Mencari nama tim yang duplikat.....                           | 57 |
| Gambar 4.14 Mengganti nama tim dengan nama baru .....                     | 58 |
| Gambar 4.15 Menggabungkan data CPI dan membuat data Adjusted Salary ..... | 58 |
| Gambar 4.16 Menghilangkan row dengan missng NaN dan data duplikat.....    | 59 |
| Gambar 4.17 Mengganti nama tim yang duplikat.....                         | 59 |
| Gambar 4.18 Menggabungkan data CPI dan membuat data Adjusted Salary ..... | 60 |
| Gambar 4.19 Melakukan Feature Selection .....                             | 61 |
| Gambar 4.20 Melakukan Grouping.....                                       | 62 |
| Gambar 4.21 Train Test Split Dataset.....                                 | 63 |
| Gambar 4.22 Membuat transformer dengan standardscaler .....               | 63 |
| Gambar 4.23 Membuat Function untuk melihat R2 dan RMSE.....               | 64 |
| Gambar 4.24 Membuat function untuk cross validation dari 5 fold .....     | 64 |
| Gambar 4.25 Membuat untuk melihat skor R2 dan RMSE dari model .....       | 65 |
| Gambar 4.26 Membuat untuk melihat skor R2 dan RMSE dari model .....       | 66 |
| Gambar 4.27 Membuat untuk melihat skor R2 dan RMSE dari model .....       | 67 |
| Gambar 4.28 Membuat pipeline model XGBoost.....                           | 69 |
| Gambar 4.29 Inisiasi model XGBoost .....                                  | 70 |
| Gambar 4.30 Hasil R2 Training dan Test model XGBoost .....                | 71 |
| Gambar 4.31 Hasil Cross validation R2 dan RMSE .....                      | 71 |
| Gambar 4.32 Hasil RMSE Training dan Test .....                            | 72 |
| Gambar 4.33 Membuat dataframe untuk menampung input .....                 | 73 |
| Gambar 4.34 Melakukan input data .....                                    | 74 |
| Gambar 4.35 Melakukan Prediksi Gaji Pemain Bisbol .....                   | 75 |
| Gambar 4.36 Input data pada aplikasi streamlit.....                       | 79 |
| Gambar 4.37 Hasil Prediksi pada Aplikasi Streamlit.....                   | 80 |

## DAFTAR LAMPIRAN

|                                |    |
|--------------------------------|----|
| Lampiran A Nama Lampiran ..... | 92 |
|--------------------------------|----|



UMN  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA