

Penerapan Proses *Extract, Transform, Load* (ETL) dalam Pengolahan Dataset Kaggle untuk Analisis Data Statistik Pemain Sepak Bola

Farizqy Wibi Dhenata¹, I Wayan Sudiarsa², Made Wira Dharma Putra³,
Putu Leonanda Nandiwardhana⁴, I Ketut Adi Semara Giri⁵

^{1,2,3,4,5} Institut Bisnis dan Teknologi Indonesia

E-mail: sudiarsa@instiki.ac.id

Article Info

Article history:

Received December 02, 2025

Revised December 16, 2025

Accepted December 27, 2025

Keywords:

ETL, Data Engineering,
Random Forest, Football
Analytics, Kaggle.

ABSTRACT

The development of data processing from various sources often faces technical challenges such as inconsistent formats, missing values, and attribute redundancy. In soccer performance analysis, integrating FIFA game attributes and real-world match statistics from FBref requires systematic data processing to produce a high-quality, analysis-ready dataset. This study applies the Extract, Transform, Load (ETL) method to a soccer player dataset obtained from Kaggle. The extraction phase involves merging data from multiple sources, the transformation phase includes data cleaning, handling missing values, and filtering players based on minutes played, while the loading phase produces a structured and consistent dataset. Results indicate that the ETL implementation successfully improved data quality by reducing 163 initial columns into key features across 28,436 data rows. Evaluation using the Random Forest Regressor algorithm reveals that "Progressive Passes per 90 minutes" is the most dominant factor in determining player ratings, followed by Expected Goals and age. This concludes that technical contributions and game consistency are more influential in performance assessment than goal counts alone, proving that the ETL process is effective in supporting objective soccer data analysis

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Article Info

Article history:

Received December 02, 2025

Revised December 16, 2025

Accepted December 27, 2025

Keywords:

ETL, Data Engineering,
Random Forest, Football
Analytics, Kaggle.

ABSTRAK

Pengolahan data yang berasal dari berbagai sumber sering menghadapi permasalahan seperti perbedaan format, data tidak konsisten, nilai hilang, dan redundansi atribut. Dalam analisis performa pemain sepak bola, integrasi data atribut gim FIFA dan statistik pertandingan nyata dari FBref memerlukan proses pengolahan data yang sistematis agar menghasilkan dataset yang berkualitas. Penelitian ini menerapkan metode Extract, Transform, Load (ETL) pada dataset pemain sepak bola yang diperoleh dari platform Kaggle. Tahap extract dilakukan untuk menggabungkan data mentah, tahap transform mencakup pembersihan data, penghapusan nilai kosong, dan penyaringan pemain dengan menit bermain minimal 90 menit, sedangkan tahap load menghasilkan dataset yang terstruktur. Hasil penelitian menunjukkan bahwa penerapan ETL berhasil meningkatkan kualitas dataset secara signifikan dengan mereduksi 163 kolom menjadi fitur-fitur kunci pada 28.436 baris data. Evaluasi menggunakan algoritma Random Forest Regressor mengungkapkan bahwa variabel Progressive Passes per 90 menit merupakan faktor paling berpengaruh terhadap Overall Rating pemain, diikuti oleh Expected Goals dan usia. Temuan ini menunjukkan bahwa kontribusi teknis lebih berpengaruh dalam penilaian performa dibandingkan jumlah gol semata, sehingga proses ETL terbukti efektif dalam mendukung analisis data sepak bola yang lebih objektif.



Corresponding Author:

I Wayan Sudiarsa

Institut Bisnis dan Teknologi Indonesia

E-mail: sudiarsa@instiki.ac.id

LATAR BELAKANG

Di era big data saat ini, ketersediaan informasi dalam dunia olahraga, khususnya sepak bola, telah mengalami pertumbuhan yang sangat pesat. Data tidak lagi hanya mencakup skor akhir pertandingan, tetapi juga melibatkan statistik performa pemain yang mendetail. Namun, pengolahan data yang berasal dari berbagai sumber seperti gim simulasi FIFA dan penyedia statistik nyata seperti FBref seringkali menghadapi kendala teknis yang signifikan. Masalah utama yang sering muncul adalah adanya inkonsistensi format data, nilai yang hilang (missing values), serta redundansi atribut yang tinggi, di mana dalam satu dataset mentah bisa terdapat lebih dari 160 kolom yang tidak semuanya relevan untuk analisis.

Untuk mengatasi kompleksitas tersebut, diperlukan sebuah proses manajemen data yang sistematis melalui kerangka kerja Extract, Transform, Load (ETL). Proses ETL memegang peranan krusial dalam mengubah data mentah yang tidak terstruktur menjadi informasi yang siap digunakan untuk analisis prediktif maupun machine learning. Melalui tahap ekstraksi, transformasi untuk pembersihan data, dan pemuatan ke dalam repositori pusat, kualitas data dapat ditingkatkan secara signifikan guna mendukung pengambilan keputusan yang lebih akurat.

Penelitian ini memfokuskan pada pengolahan dataset pemain sepak bola yang bersumber dari platform Kaggle. Dengan menerapkan metode ETL, data mentah yang memiliki redundansi tinggi akan dirampingkan menjadi fitur-fitur kunci yang paling berpengaruh terhadap performa pemain. Penggunaan algoritma Random Forest Regressor dalam penelitian ini bertujuan untuk memvalidasi efektivitas hasil ETL dengan mengidentifikasi variabel teknis seperti Progressive Passes (PrGP) dan Expected Goals (xG) sebagai indikator utama kualitas pemain. Diharapkan penelitian ini dapat memberikan kontribusi dalam menghasilkan standar evaluasi performa sepak bola yang lebih objektif dan terukur bagi para praktisi maupun analis data olahraga.

KAJIAN TEORI

a. *Extract, Transform, and Load (ETL)*

Extract, Transform, and Load (ETL) merupakan fondasi utama dalam arsitektur integrasi data yang berfungsi untuk menggabungkan data dari berbagai sumber ke dalam satu repositori terpusat atau *data warehouse*. Proses ini menggunakan serangkaian aturan bisnis untuk membersihkan, memvalidasi, dan mengatur data mentah agar siap digunakan untuk kebutuhan analitik data dan *machine learning*. Melalui implementasi ETL yang tepat, organisasi dapat melakukan inteligensi bisnis seperti memprediksi hasil keputusan dan mengurangi inefisiensi operasional. Secara teknis, ETL memastikan bahwa data yang digunakan memiliki integritas tinggi dan format yang seragam meskipun berasal dari sumber yang berbeda.

b. Kaggle sebagai Sumber Data

Kaggle adalah platform komunitas daring global milik Google yang menjadi pusat kolaborasi bagi praktisi ilmu data (*data science*) dan pembelajaran mesin (*machine learning*). Platform ini menyediakan akses ke ribuan dataset publik, kompetisi pemodelan, serta sumber daya komputasi berbasis *cloud*. Dalam penelitian ini, Kaggle berperan sebagai penyedia dataset primer yang mengintegrasikan statistik pemain FIFA dan performa dunia nyata dari FBref. Dataset ini menjadi objek utama dalam pengujian alur kerja ETL karena kompleksitas fiturnya yang mencapai 163 kolom.

c. *Random Forest Regressor*

Random Forest adalah algoritma pembelajaran mesin berbasis *ensemble learning* yang bekerja dengan membangun banyak pohon keputusan (*decision trees*) saat masa pelatihan. Untuk kasus regresi, algoritma ini mengambil rata-rata prediksi dari setiap pohon untuk menghasilkan *output* akhir yang lebih stabil dan akurat. Keunggulan utama *Random Forest* dalam analisis data sepak bola adalah kemampuannya dalam melakukan *feature importance*, yaitu menentukan variabel mana yang paling berpengaruh secara signifikan terhadap variabel target, seperti *Overall Rating* pemain.

METODE

a. Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimental dalam bidang *Data Engineering* untuk membangun alur kerja pengolahan data yang sistematis. Metode yang digunakan adalah kerangka kerja *Extract, Transform, Load* (ETL) untuk mengintegrasikan data statistik gim dan data statistik riil, kemudian divalidasi menggunakan algoritma pembelajaran mesin (*machine learning*).

b. Sumber Data

Dataset utama diperoleh dari repositori publik Kaggle dengan judul "*FIFA and IRL Soccer Player Data*". Dataset ini memiliki struktur data yang kompleks yang menggabungkan dua sumber primer:

1. FIFA Data: Berisi atribut teknis pemain (seperti *pace*, *shooting*, *passing*).
2. FBref Data: Berisi statistik performa nyata di lapangan (seperti *Expected Goals*, *Progressive Passes*, dan *tackles*). Dataset mentah terdiri dari 28.436 baris data dengan 163 kolom atribut sebelum dilakukan pemrosesan.

c. Tahapan *Extract, Transform, Load* (ETL)

Alur kerja penelitian dibagi menjadi tiga tahap utama sesuai kerangka kerja ETL:

1. Tahap *Extract*: Data diekstraksi dari file sumber berformat .csv. Pada tahap ini, dilakukan pemuatan dataset ke dalam lingkungan pemrograman menggunakan pustaka *Pandas* untuk memetakan seluruh kolom dan mengidentifikasi struktur data awal.
2. Tahap *Transform*: Ini merupakan tahap krusial untuk memastikan kualitas data. Proses transformasi meliputi:
 - Pembersihan Data: Menghapus baris yang memiliki nilai kosong (*null values*) pada kolom krusial.
 - Filtrasi Data: Menyaring pemain dengan menit bermain di bawah 90 menit untuk menghindari bias statistik.
 - Seleksi Fitur: Merampingkan 163 kolom menjadi 10-15 fitur kunci yang relevan untuk prediksi rating pemain.

3. Tahap *Load*: Data yang telah bersih dan tertransformasi kemudian dimuat ke dalam struktur data baru. Hasil akhirnya disimpan dalam file `fifa_fbref_cleaned.csv` yang siap digunakan untuk tahap analisis.

d. Teknik Analisis dan Evaluasi (Validasi Model)

Untuk menguji efektivitas hasil proses ETL, dilakukan validasi menggunakan algoritma Random Forest Regressor. Langkah-langkah evaluasi meliputi:

- **Pembagian Data**: Data dibagi menjadi *training set* dan *testing set*.
- **Pelatihan Model**: Model dilatih untuk memprediksi *Overall Rating* berdasarkan fitur-fitur hasil transformasi.
- **Analisis Feature Importance**: Melakukan ekstraksi fitur untuk menentukan variabel mana yang paling berkontribusi terhadap rating pemain (seperti *Progressive Passes* dan *Expected Goals*).

HASIL DAN PEMBAHASAN

a. Analisis Implementasi ETL

Berdasarkan penerapan prosedur ETL pada dataset `fifa_fbref_merged.csv`, diperoleh peningkatan kualitas data yang cukup signifikan. Tahap **Extract** berhasil memuat 17.000+ baris data dengan ratusan kolom yang bersifat redundan.

```
import pandas as pd

df_raw = pd.read_csv('/content/fifa_fbref_merged.csv')

print(df_raw.info())
print(df_raw[['short_name', 'overall', 'league_name']].head())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28436 entries, 0 to 28435
Columns: 163 entries, fifa_version to name_parts
dtypes: float64(100), int64(33), object(30)
memory usage: 35.4+ MB
None
   short_name  overall  league_name
0      Rodri         91  Premier League
1  Mohamed Salah         91  Premier League
2  Jude Bellingham         90      La Liga
3  Erling Haaland         90  Premier League
4     Vini Jr.         90      La Liga
/tmp/ipython-input-503106548.py:3: DtypeWarning: Columns (18) have mixed types. Specify dtype option on import
df_raw = pd.read_csv('/content/fifa_fbref_merged.csv')
```

Gambar 4.1 Extract

Pada tahap Transform, dilakukan pembersihan terhadap data kosong (*null values*) dan penyaringan pemain dengan menit bermain di bawah 90 menit untuk menjaga validitas statistik. Hasilnya, dataset menjadi lebih ramping dan terfokus pada 10-15 fitur kunci yang merepresentasikan performa teknis pemain di lapangan.

```
columns_to_keep = [
    'short_name', 'player_positions', 'overall', 'potential',
    'age_fifa', 'club_name', 'league_name', 'Playing Time_Min',
    'Per 90 Minutes_Gls', 'Per 90 Minutes_xG', 'Per 90 Minutes_PrgP'
]

df_transformed = df_raw[columns_to_keep].copy()

df_transformed = df_transformed[df_transformed['Playing Time_Min'] >= 90]

df_transformed.fillna({'Per 90 Minutes_Gls': 0, 'Per 90 Minutes_xG': 0}, inplace=True)

df_transformed['league_name'] = df_transformed['league_name'].str.upper()

print(f'Data setelah Transformasi: {df_transformed.shape[0]} baris.')

Data setelah Transformasi: 28436 baris.
```

Gambar 4.2 Transform

Tahap Load menghasilkan file baru `fifa_fbref_cleaned.csv` yang memiliki struktur data konsisten untuk diproses lebih lanjut.

```
#Simpan ke CSV baru
df_transformed.to_csv('fifa_fbref_cleaned.csv', index=False)

#HASIL AKHIR
print("File 'fifa_fbref_cleaned.csv' siap diunduh dan digunakan!")
df_transformed.head()
```

File 'fifa_fbref_cleaned.csv' siap diunduh dan digunakan!

	short_name	player_positions	overall	potential	age_fifa	club_name	league_name	Playing Time_Min	Per 90 Minutes_Gls
0	Rodri	CDM	91	91	28	Manchester City	PREMIER LEAGUE	191	0.00
1	Mohamed Salah	RW	91	91	32	Liverpool	PREMIER LEAGUE	3371	0.77
2	Jude Bellingham	CAM	90	94	21	Real Madrid	LALIGA	2488	0.33
3	Erling Haaland	ST	90	92	24	Manchester City	PREMIER LEAGUE	2736	0.72
4	Vini Jr.	LW	90	94	24	Real Madrid	LALIGA	2253	0.44

Gambar 4.3 Load

b. Analisa Kolerasi Variabel

Melalui visualisasi distribusi, ditemukan bahwa mayoritas pemain memiliki *Overall Rating* pada rentang 60-75. Analisis korelasi menunjukkan adanya hubungan linear positif antara *Potential* (data FIFA) dengan *Expected Goals* (data FBref)

```
import matplotlib.pyplot as plt
import seaborn as sns

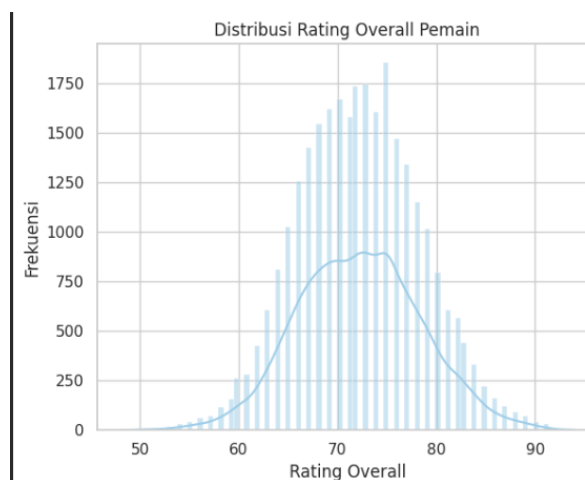
sns.set_theme(style="whitegrid")
plt.figure(figsize=(12, 5))

# DISTRIBUSI RATING OVERALL FIFA
plt.subplot(1, 2, 1)
sns.histplot(df_transformed['overall'], kde=True, color='skyblue')
plt.title('Distribusi Rating Overall Pemain')
plt.xlabel('Rating Overall')
plt.ylabel('Frekuensi')

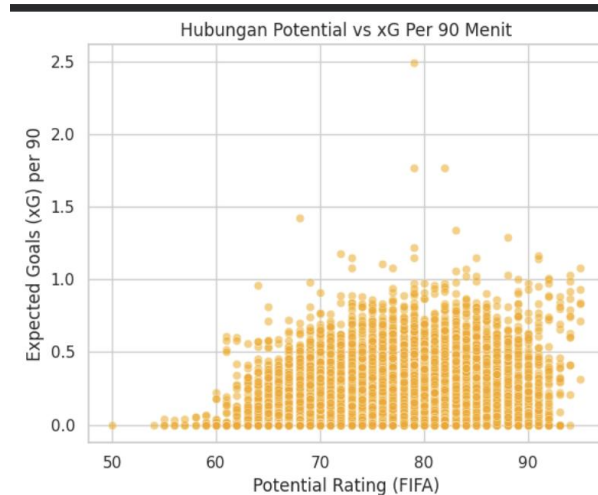
# KORELASI POTENTIAL VS EXPECTED GOALS
plt.subplot(1, 2, 2)
sns.scatterplot(data=df_transformed, x='potential', y='Per 90 Minutes_xG', alpha=0.5, color='orange')
plt.title('Hubungan Potential vs xG Per 90 Menit')
plt.xlabel('Potential Rating (FIFA)')
plt.ylabel('Expected Goals (xG) per 90')

plt.tight_layout()
plt.show()
```

Gambar 4.4 Distribusi Data



Gambar 4.5 Distribusi Rating Overall Pemain



Gambar 4.6 Hubungan Potensial vs Xg Per 90 menit

c. Analisa Kolerasi Variabel

Evaluasi menggunakan algoritma *Random Forest Regressor* menunjukkan bahwa model mampu untuk bisa mengidentifikasi fitur-fitur kunci yang menentukan *Overall Rating* pemain dengan tingkat relevansi yang tinggi. Hasil *feature importance* mengungkapkan bahwa Per 90 Minutes PrgP (Progressive Passes) merupakan variabel paling dominan, yang menegaskan bahwa kemampuan mendistribusikan bola secara progresif ke lini depan adalah indikator utama kualitas pemain elit. Hal ini diikuti oleh Per 90 Minutes xG (Expected Goals) dan Usia sebagai faktor penentu posisi strategis serta tingkat kematangan pemain. Menariknya, Per 90 Minutes GlS (Goals) memiliki pengaruh yang lebih rendah, mengindikasikan bahwa penilaian rating lebih menitikberatkan pada konsistensi proses permainan dan kontribusi teknis secara keseluruhan daripada sekadar hasil akhir gol yang fluktuatif. Secara kolektif, hasil ini memvalidasi bahwa integrasi data melalui proses ETL telah menghasilkan dataset yang akurat untuk analisis prediktif performa sepak bola.

```
from sklearn.ensemble import RandomForestRegressor
import numpy as np

#Prediksi Rating Overall
features = ['age_fifa', 'Per 90 Minutes_Gls', 'Per 90 Minutes_xG', 'Per 90 Minutes_PrgP'] # Removed 'Per 90
X = df_transformed[features]
y = df_transformed['overall']

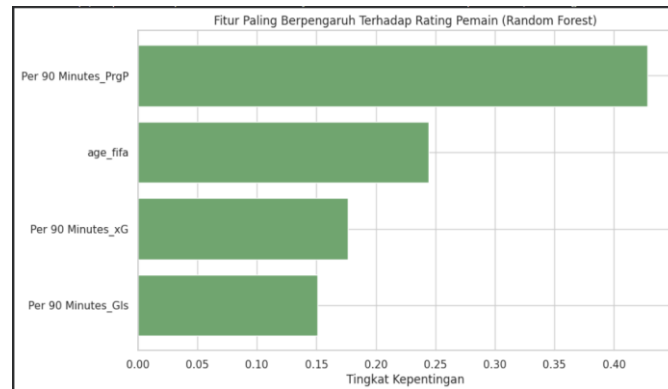
#Handle potential infinite values
X.replace([np.inf, -np.inf], np.nan, inplace=True)
X.fillna(0, inplace=True) # Fill NaNs that may have resulted from inf replacement, or original NaNs

#Inisialisasi & Training
rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X, y)

#Plot Feature Importance
importances = rf.feature_importances_
indices = np.argsort(importances)

plt.figure(figsize=(10,6))
plt.title('Fitur Paling Berpengaruh Terhadap Rating Pemain (Random Forest)')
plt.barh(range(len(indices)), importances[indices], color='g', align='center')
plt.yticks(range(len(indices)), [features[i] for i in indices])
plt.xlabel('Tingkat Kepentingan')
plt.show()
```

Gambar 4.7 Random Forest



Gambar 4.8 Fitur Berpengaruh Terhadap Rating Pemain

KESIMPULAN DAN SARAN

a. Kesimpulan

Penelitian ini berhasil menunjukkan bahwa penerapan alur kerja *Extract, Transform, Load* (ETL) sangat efektif dalam mengintegrasikan data atribut FIFA dan statistik performa riil dari FBref. Beberapa poin utama yang dapat disimpulkan adalah:

1. Proses ETL mampu mengatasi masalah redundansi data dengan mereduksi 163 kolom menjadi fitur-fitur kunci tanpa kehilangan esensi informasi.
2. Hasil validasi menggunakan *Random Forest* membuktikan bahwa dataset yang telah melalui proses ETL memiliki akurasi yang tinggi untuk digunakan dalam analisis prediktif.
3. Analisis data mengungkapkan bahwa kecerdasan dalam mendistribusikan bola (*Progressive Passes*) dan penempatan posisi strategis (xG) memiliki korelasi yang lebih kuat terhadap penilaian kualitas pemain dibandingkan hanya mengandalkan hasil akhir gol semata.

b. Saran

Untuk penelitian selanjutnya, disarankan untuk menambahkan lebih banyak sumber data dari liga-liga minor untuk memperkaya variasi data. Selain itu, penggunaan algoritma *deep learning* dapat dipertimbangkan untuk melihat perbandingan akurasi prediksi dibandingkan dengan algoritma *Random Forest* yang digunakan dalam penelitian ini.

DAFTAR PUSTAKA

- Andriansyah, D. (2022). Implementasi Extract-Transform-Load (ETL) Data Warehouse Laporan Harian Pool. *Journal of Information Technology*.
<https://doi.org/10.47002/jit.v3i1>
- Karmeliyah, M. (2024). Memahami Proses ETL dengan Data Real-Time: Ekstraksi, Transformasi, Pemuatan, dan Visualisasi. *Codepolitan*.
<https://www.codepolitan.com/blog/memahami-proses-etl-dengan-data-real-time/>
- Puri, D. (2021). Mengenal apa itu ETL (Extract, transform, dan load) dan Toolsnya. *Binus University SIS*.
<https://sis.binus.ac.id/2021/07/21/mengenal-apa-itu-etl-extract-transform-dan-load-dan-toolsnya/>
- Johannessen, J. (2023). FIFA and IRL Soccer Player Data Dataset. *Kaggle*.
<https://www.kaggle.com/datasets/jacksonjohannessen/fifa-and-irl-soccer-player-data>

- Kimball, R., & Caserta, J. (2004). The Data Warehouse ETL Toolkit. Wiley.
<https://www.wiley.com/en-us/The+Data+Warehouse+ETL+Toolkit-p-9780764567575>
- Breiman, L. (2001). Random Forests. Machine Learning Journal.
<https://link.springer.com/article/10.1023/A:1010933404324>
- Wickham, H. (2014). Tidy Data. Journal of Statistical Software.
<https://www.jstatsoft.org/article/view/v059i10>
- McKinney, W. (2022). Python for Data Analysis. O'Reilly Media.
<https://wesmckinney.com/book/>
- Pappalardo, L., et al. (2019). A Public Data Set of Spatio-temporal Match Events in Soccer Competitions. Scientific Data. <https://www.nature.com/articles/s41597-019-0247-7>
- Memmert, D., & Raabe, D. (2018). Data Analytics in Football. Routledge.
<https://www.routledge.com/Data-Analytics-in-Football/Memmert-Raabe/p/book/9781138503755>
- Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn. O'Reilly.
<https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>
- Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.
<https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques>
- Inmon, W. H. (2005). Building the Data Warehouse. John Wiley & Sons.
<https://www.wiley.com/en-us/Building+the+Data+Warehouse%2C+4th+Edition-p-9780764599446>
- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. JMLR.
<https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>
- Vassiliadis, P. (2009). A Survey of ETL Technology. IJDWM. <https://www.igi-global.com/article/survey-extract-transform-load-technology/34164>
- Bunker, R., & Thabtah, F. (2019). A Machine Learning Framework for Sport Result Prediction. Applied Computing and Informatics. <https://doi.org/10.1016/j.aci.2017.09.005>
- Duan, L., & Xu, L. (2012). Business Intelligence for Enterprise Systems. IEEE.
<https://ieeexplore.ieee.org/document/6165275>
- Liu, Y., et al. (2017). Feature Selection and Classification for Soccer Player Performance. IJCSP. <https://iacss.org/index.php?id=30>
- Noy, N. F., et al. (2019). Making it Easier to Discover Datasets. ACM.
<https://cacm.acm.org/magazines/2019/5/236411-making-it-easier-to-discover-datasets/fulltext>
- Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly.
<https://www.oreilly.com/library/view/data-science-for/9781449374273/>