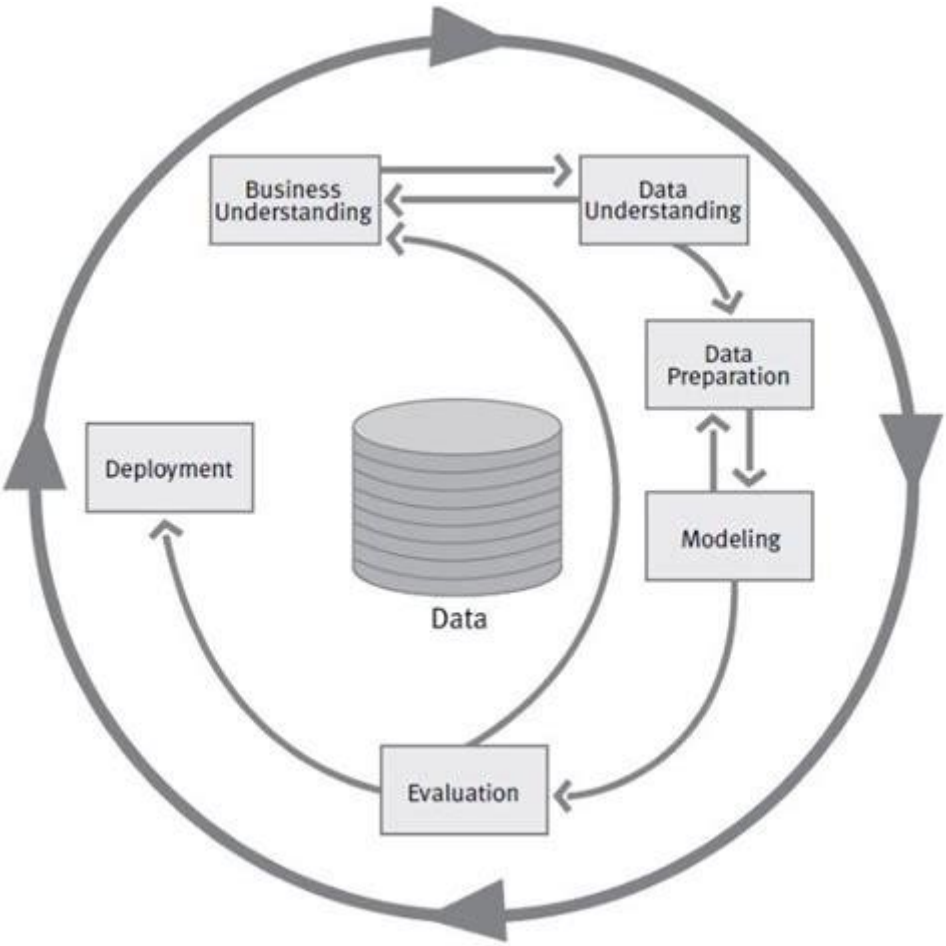


1. Data Understanding

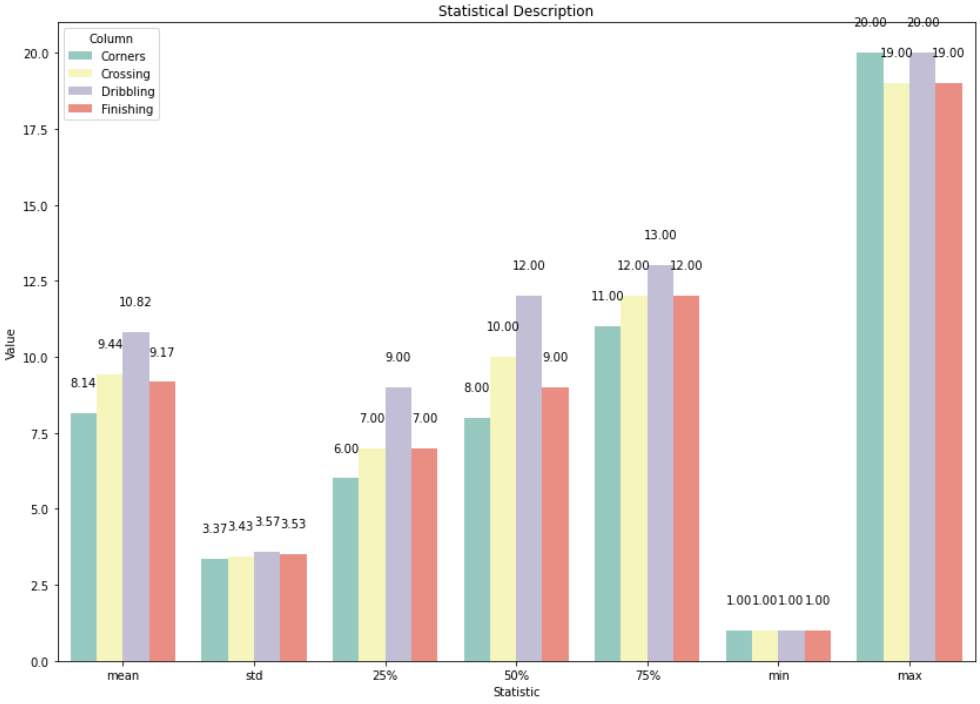
Tugas-tugas penting dalam tahap ini meliputi pengumpulan data dari berbagai sumber, eksplorasi yang cermat, deskripsi yang teliti, dan evaluasi yang ketat terhadap kualitas data. Untuk menjelaskan proses ini lebih lanjut, panduan pengguna merinci tugas deskripsi data melalui penerapan analisis statistik untuk memastikan atribut-atribut dan korelasinya.

a. Analisis Univariat

Analisis univariat didefinisikan sebagai analisis yang dilakukan hanya pada satu ("uni") variabel ("variate") untuk merangkum atau menggambarkan variabel tersebut. Dalam analisis univariat, biasanya teknik statistik deskriptif digunakan untuk merangkum dan memahami karakteristik serta distribusi dari satu variabel tunggal. Ini melibatkan penggunaan ukuran seperti rata-rata, median, modus, rentang, varians, dan deviasi standar, serta representasi grafis seperti histogram, box plot, atau diagram batang. Tujuannya adalah untuk memperoleh

pemahaman mengenai perilaku dan karakteristik variabel tertentu tanpa mempertimbangkan hubungannya dengan variabel lain.

Dilakukan analisis univariat kepada variabel independen dan dependen yang telah dipilih. Variabel independen terdiri dari variabel: "Corners", "Crossing", "Dribbling", "Finishing", "First Touch", "Free Kick Taking", "Heading", "Long Shots", "Passing", "Tackling", "Technique", "Concentration", "Vision", "Decision", "Determination", "Position.1", "Teamwork", "Balance", "Natural Fitness", "Pace", "Stamina", "Strength", “Left Foot”, “Right Foot.” Variabel dependen terdiri dari: "DL", "DC", "DR", "WBL", "WBR", "DM", "MR", "ML", "MC", "AML", "AMC", "AMR", "ST."



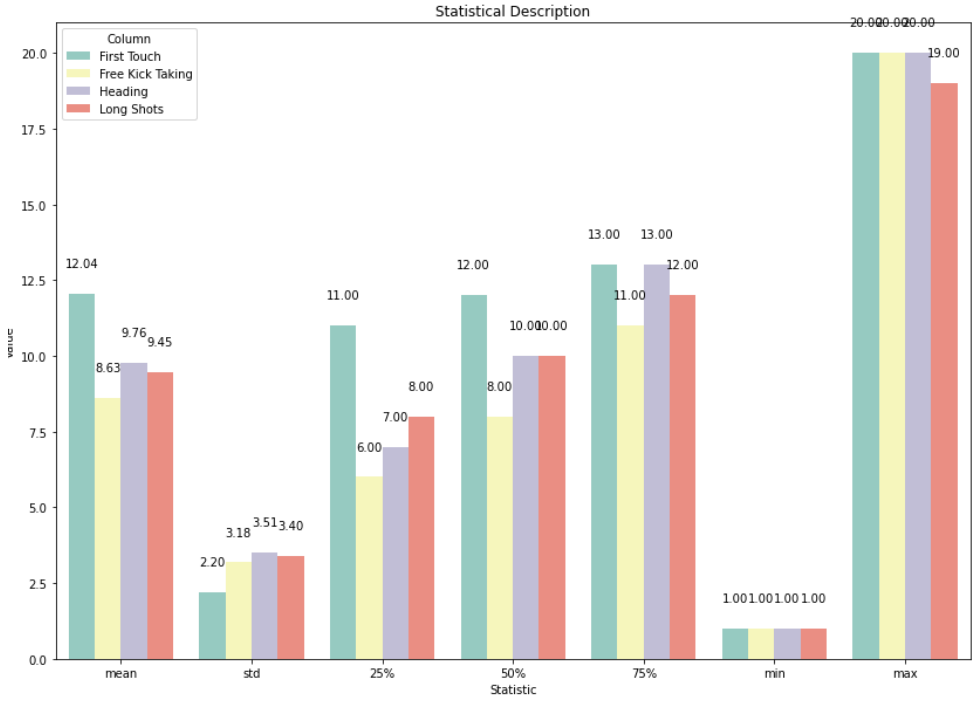
Gambar 1. 1 Diagram Batang Statistik Deskriptif dari Variabel *Corners, Crossing, Dribbling,* dan *Finishing.*

Gambar 1.1. merupakan statistik deskriptif dari variabel *Corners, Crossing, Dribbling*, dan *Finishing*. Statistik deskriptif pada Gambar 1.1 menunjukkan statistik *mean*, standar deviasi, kuartil pertama (25%), kuartil kedua / median (50%), kuartil ketiga (75%), nilai minimum, dan maksimum. Tabel 1.1 merupakan rangkuman dari hasil visualisasi pada Gambar 1.1. dalam bentuk tabular.

Tabel 1. 1 Tabel Statistik Deskriptif dari Variabel *Corners, Crossing, Dribbling,* dan *Finishing.*

Variabel	**Standar Deviasi**	***Mean***	**Kuartil pertama**	**Kuartil kedua / Median**	**Kuartil ketiga**	**Minimum**	**Maksimum**
Corners	3.37	8.14	6	8	11	1	20
Crossing	3.43	9.44	7	10	12	1	19
Dribbling	3.57	10.82	9	12	13	1	20
Finishing	3.53	9.17	7	9	12	1	19

Dapat dihasilkan asumsi terhadap data deskriptif dari Gambar 1.1 dan Tabel 1.1. Keahlian *Corners* merupakan keahlian yang paling sedikit dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terkecil dibandingkan dengan *Crossing*, *Dribbling*, dan *Finishing*. Keahlian *Dribbling* di sisi lain merupakan keahlian yang paling dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga paling besar di antara keahlian *Crossing*, *Corners*, dan *Finishing*.



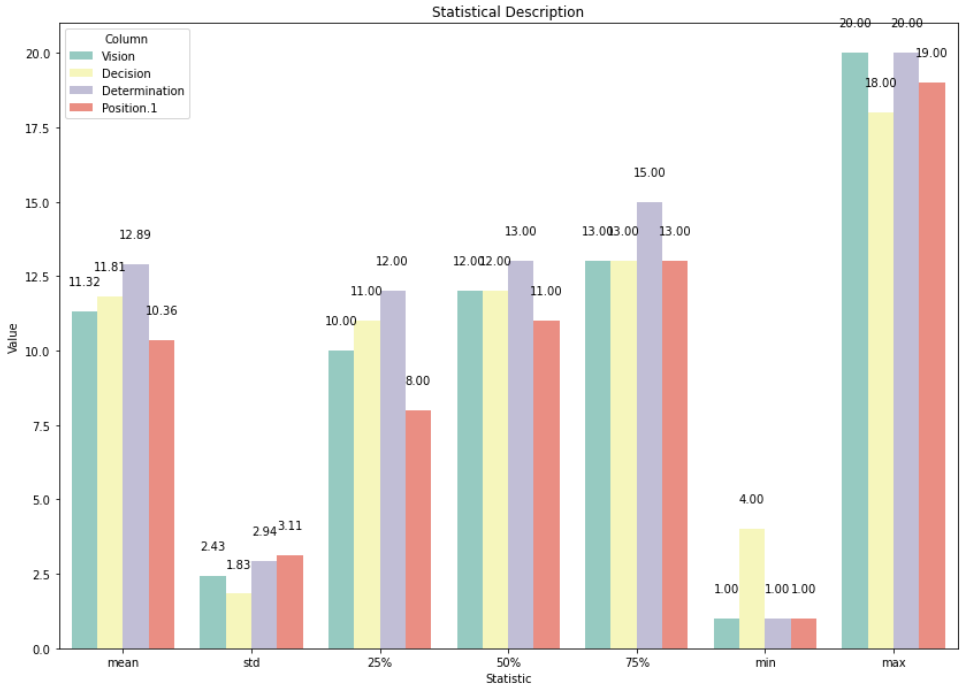
Gambar 1. 2 Diagram Batang Statistik Deskriptif dari Variabel *First Touch, Free Kick Taking, Heading,* dan *Long Shots.*

Gambar 1.2. merupakan statistik deskriptif dari variabel *First Touch, Free Kick Taking, Heading*, dan *Long Shots*. Statistik deskriptif pada Gambar 1. 2 menunjukkan statistik *mean*, standar deviasi, kuartil pertama (25%), kuartil kedua / median (50%), kuartil ketiga (75%), nilai minimum, dan maksimum. Tabel 1.2 merupakan rangkuman dari hasil visualisasi pada Gambar 1.2. dalam bentuk tabular.

Tabel 1. 2 Tabel Statistik Deskriptif dari Variabel *First Touch, Free Kick Taking, Heading,* dan *Long Shots*

Variabel	**Standar Deviasi**	***Mean***	**Kuartil pertama**	**Kuartil kedua / Median**	**Kuartil ketiga**	**Minimum**	**Maksimum**
First Touch	2.2	12.04	11	12	13	1	20
Free Kick Taking	3.18	8.63	6	8	11	1	20
Heading	3.51	9.76	7	10	13	1	20
Long Shots	3.4	9.45	8	10	12	1	19

Dapat dihasilkan asumsi terhadap data deskriptif dari Gambar 1.2 dan Tabel 1.2. Keahlian *Free Kick Taking* merupakan keahlian yang paling sedikit dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terkecil dibandingkan dengan *First Touch, Heading*, dan *Long Shots*. Keahlian *First Touch* di sisi lain merupakan keahlian yang paling dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga paling besar di antara keahlian *Free Kick Taking, Heading*, dan *Long Shots*.



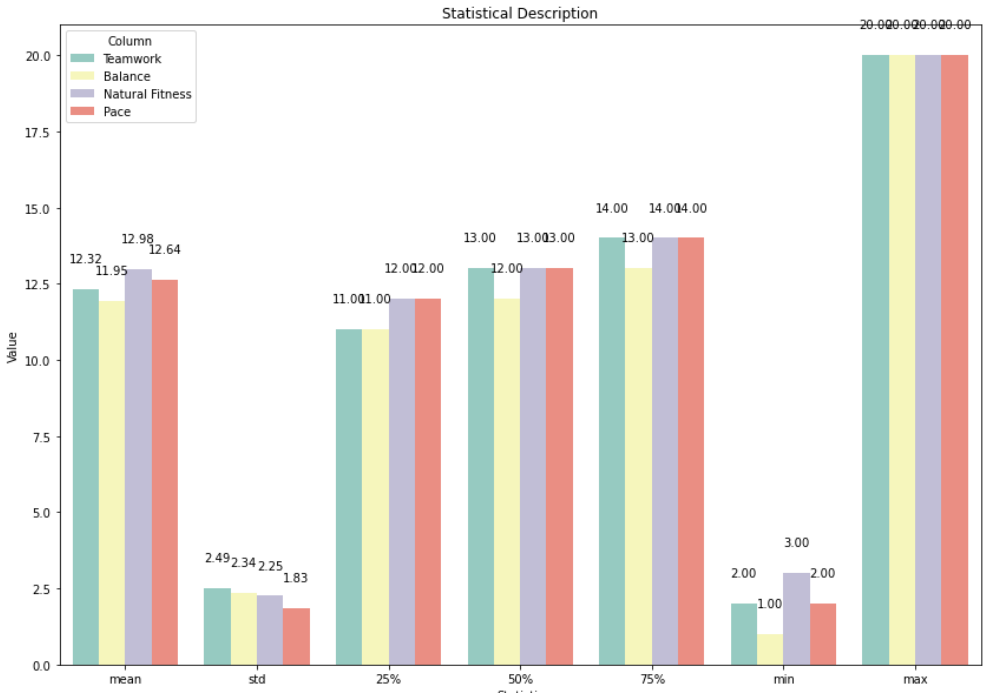
Gambar 1. 3 Diagram Batang Statistik Deskriptif dari Variabel *Vision, Decision, Determination,* dan *Positioning* (Position.1)

Gambar 1.3. merupakan statistik deskriptif dari variabel *Vision, Decision, Determination,* dan *Positioning* (Position.1). Statistik deskriptif pada Gambar 1. 3 menunjukkan statistik *mean*, standar deviasi, kuartil pertama (25%), kuartil kedua / median (50%), kuartil ketiga (75%), nilai minimum, dan maksimum. Tabel 1.3 merupakan rangkuman dari hasil visualisasi pada Gambar 1.3. dalam bentuk tabular.

Tabel 1. 3 Tabel Statistik Deskriptif dari Variabel Vision, Decision, Determination, *dan* Positioning *(Position.1)*

Variabel	**Standar Deviasi**	***Mean***	**Kuartil pertama**	**Kuartil kedua / Median**	**Kuartil ketiga**	**Minimum**	**Maksimum**
Vision	2.43	11.32	10	12	13	1	20
Decision	1.83	11.81	11	12	13	4	18
Determination	2.94	12.89	12	13	15	1	20
Positioning (Position.1)	3.11	10.36	8	11	13	1	19

Dapat dihasilkan asumsi terhadap data deskriptif dari Gambar 1.3 dan Tabel 1.3. Keahlian *Positioning* merupakan keahlian yang paling sedikit dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terkecil dibandingkan dengan *Vision, Decision*, dan *Determination*. Keahlian *Determination* di sisi lain merupakan keahlian yang paling dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga paling besar di antara keahlian *Decision, Vision*, dan *Positioning*.



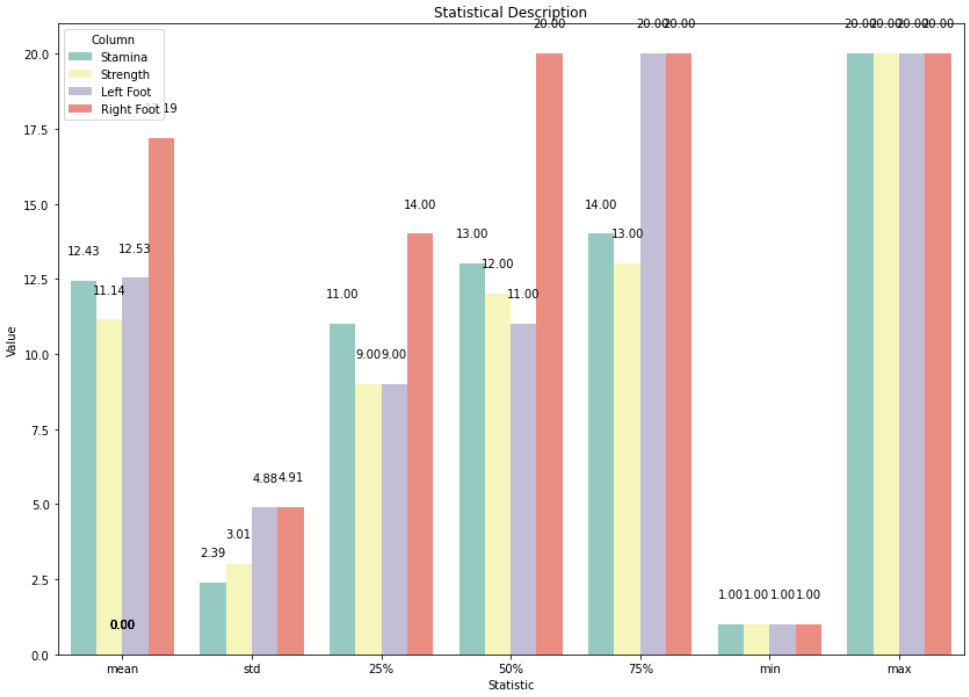
Gambar 1. 4 Diagram Batang Statistik Deskriptif dari Variabel *Teamwork, Balance, Natural Fitness,* dan *Pace*.

Gambar 1.4. merupakan statistik deskriptif dari variabel *Teamwork, Balance, Natural Fitness,* dan *Pace*. Statistik deskriptif pada Gambar 1. 4 menunjukkan statistik *mean*, standar deviasi, kuartil pertama (25%), kuartil kedua / median (50%), kuartil ketiga (75%), nilai minimum, dan maksimum. Tabel 1.4 merupakan rangkuman dari hasil visualisasi pada Gambar 1.4. dalam bentuk tabular.

Tabel 1. 4 Tabel Statistik Deskriptif dari Variabel *Teamwork, Balance, Natural Fitness,* dan *Pace*

Variabel	**Standar Deviasi**	***Mean***	**Kuartil pertama**	**Kuartil kedua / Median**	**Kuartil ketiga**	**Minimum**	**Maksimum**
Teamwork	2.49	12.32	11	13	14	2	20
Balance	2.34	11.95	11	13	14	1	20
Natural Fitness	2.25	12.98	12	12	13	3	20
Pace	1.83	12.64	12	13	14	2	20

Dapat dihasilkan asumsi terhadap data deskriptif dari Gambar 1.4 dan Tabel 1.4. Keahlian *Balance* merupakan keahlian yang paling sedikit dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terkecil dibandingkan dengan *Teamwork, Natural Fitness*, dan *Pace*. Keahlian *Natural Fitness* di sisi lain merupakan keahlian yang paling dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga paling besar di antara keahlian *Teamwork, Balance*, dan *Pace*.



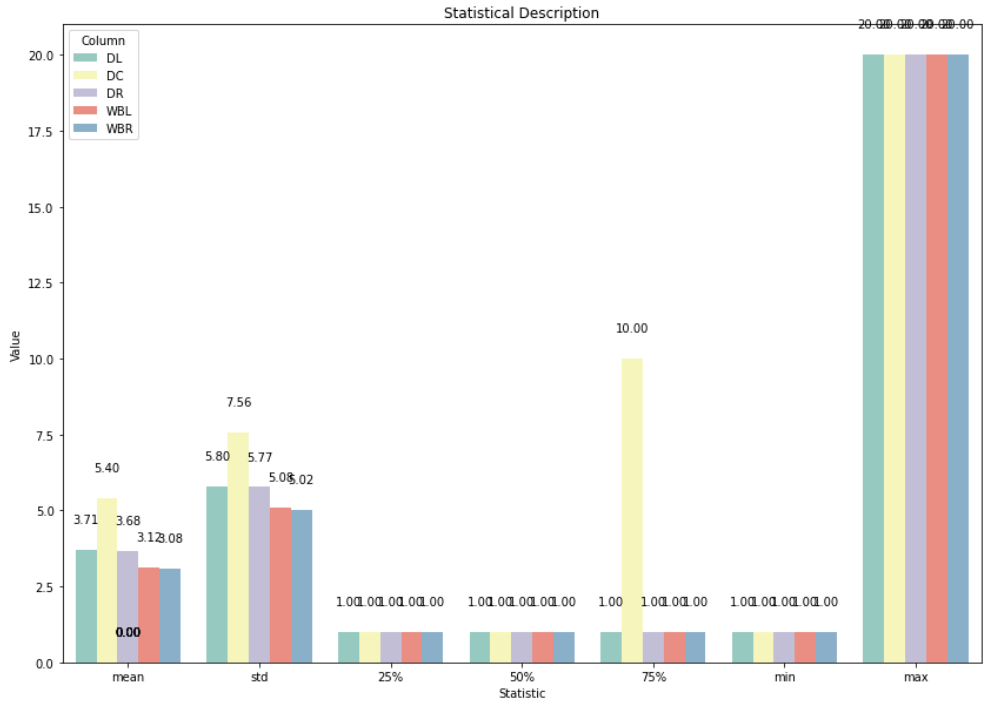
Gambar 1. 5 Diagram Batang Statistik Deskriptif dari Variabel *Stamina, Strength, Left Foot,* dan *Right Foot*.

Gambar 1.5. merupakan statistik deskriptif dari variabel *Stamina, Strength, Left Foot,* dan *Right Foot*. Statistik deskriptif pada Gambar 1. 5 menunjukkan statistik *mean*, standar deviasi, kuartil pertama (25%), kuartil kedua / median (50%), kuartil ketiga (75%), nilai minimum, dan maksimum. Tabel 1.5 merupakan rangkuman dari hasil visualisasi pada Gambar 1.5. dalam bentuk tabular.

Tabel 1. 5 Tabel Statistik Deskriptif dari Variabel *Stamina, Strength, Left Foot,* dan *Right Foot*

Variabel	Standar Deviasi	*Mean*	Kuartil pertama	Kuartil kedua / Median	Kuartil ketiga	Minimum	Maksimum
Stamina	2.39	12.43	11	13	14	1	20
Strength	3	11.14	9	12	13	1	20
Left Foot	4.88	12.53	12	11	20	1	20
Right Foot	4.91	17.9	14	20	20	1	20

Dapat dihasilkan asumsi terhadap data deskriptif dari Gambar 1.5 dan Tabel 1.5. Keahlian *Strength* merupakan keahlian yang paling sedikit dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terkecil dibandingkan dengan *Stamina*. Mayoritas pemain sepakbola memiliki kaki dominan kaki kanan dibandingkan kaki kiri, terlihat dari rata-ratanya dimana kaki kanan sebesar 17.9, sedangkan kaki kiri sebesar 12.53.



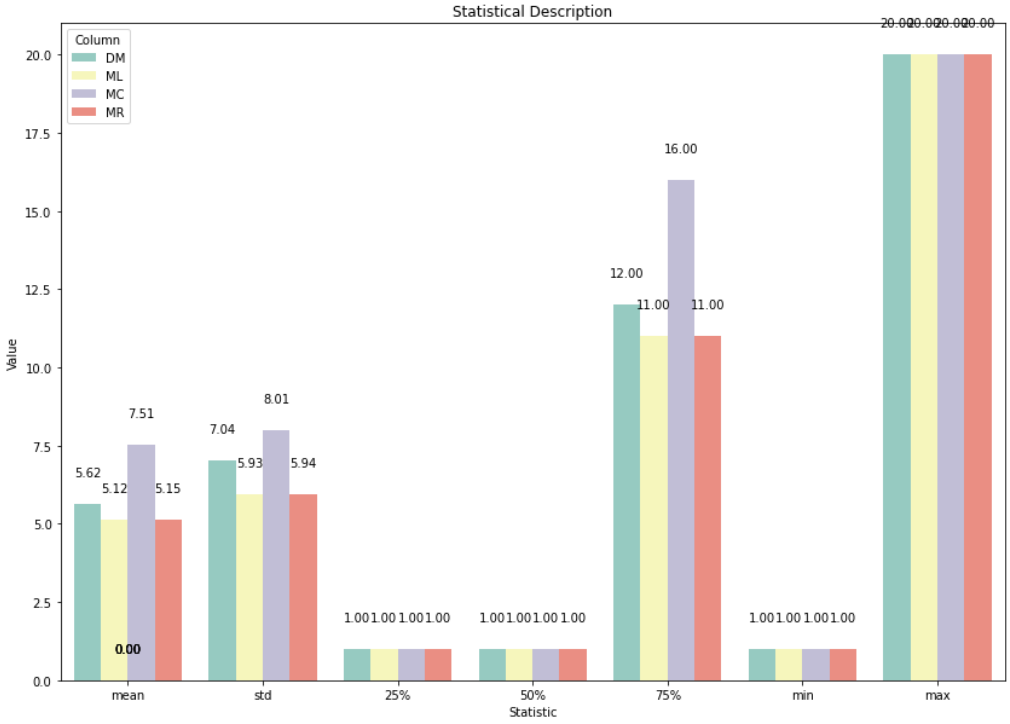
Gambar 1. 6 Diagram Batang Statistik Deskriptif dari Variabel Dependen *DL, DC, DR, WBL,* dan *WBR*.

Gambar 1.6. merupakan statistik deskriptif dari variabel dependen *DL, DC, DR,* dan *WBL,* dan *WBR*. Statistik deskriptif pada Gambar 1. 6 menunjukkan statistik *mean*, standar deviasi, kuartil pertama (25%), kuartil kedua / median (50%), kuartil ketiga (75%), nilai minimum, dan maksimum. Tabel 1.6 merupakan rangkuman dari hasil visualisasi pada Gambar 1.6. dalam bentuk tabular.

Tabel 1. 6 Tabel Statistik Deskriptif dari Variabel Dependen *DL, DC, DR, WBL,* dan *WBR*.

Variabel	**Standar Deviasi**	***Mean***	**Kuartil pertama**	**Kuartil kedua / Median**	**Kuartil ketiga**	**Minimum**	**Maksimum**
DL	5.8	3.71	1	1	1	1	20
DC	7.56	5.4	1	1	10	1	20
DR	5.77	3.68	1	1	1	1	20
WBL	5.08	3.12	1	1	1	1	20
WBR	5.02	2.08	1	1	1	1	20

Dapat dihasilkan asumsi terhadap data deskriptif dari Gambar 1.6 dan Tabel 1.6. Posisi bertahan *WBR* (*Wing Back Right*) merupakan posisi yang paling sedikit dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terkecil dibandingkan dengan posisi bertahan lainnya. Di sisi lain, posisi bertahan *DC* (*Defender Center*) merupakan posisi yang paling banyak dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terbesar dibandingkan dengan posisi bertahan lainnya. Dari data juga dapat diasumsikan bahwa banyak dari pemain tidak memiliki keahlian sama sekali pada banyak posisi bertahan, dibuktikan dengan banyak data bernilai 1.



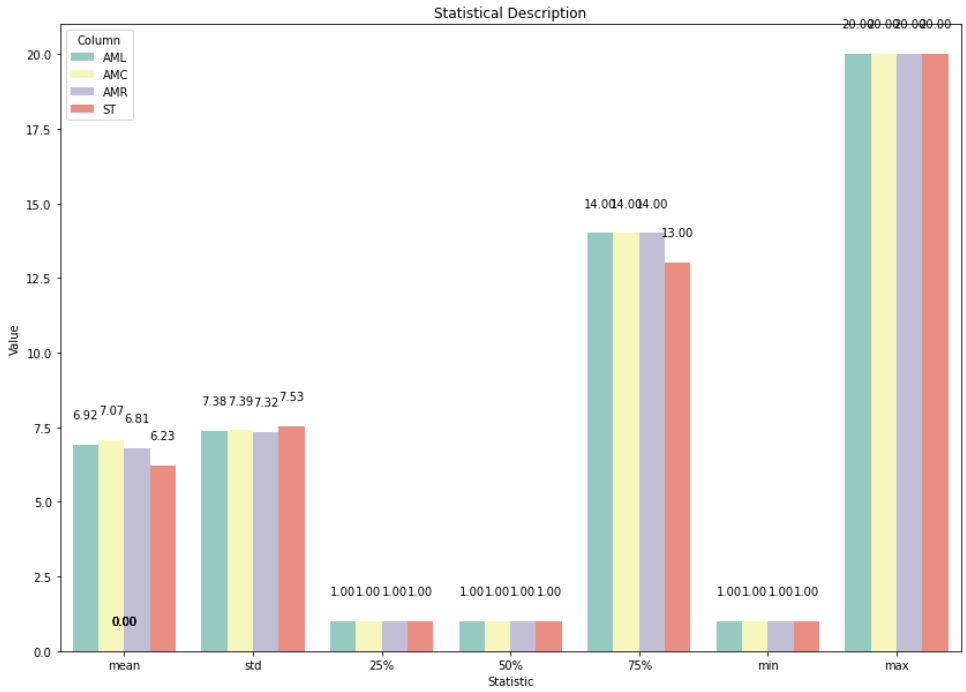
Gambar 1. 7 Diagram Batang Statistik Deskriptif dari Variabel Dependen *DM, ML, MC,* dan *MR*

Gambar 1.7. merupakan statistik deskriptif dari variabel dependen *DM, ML, MC,* dan dan *MR*. Statistik deskriptif pada Gambar 1. 7 menunjukkan statistik *mean*, standar deviasi, kuartil pertama (25%), kuartil kedua / median (50%), kuartil ketiga (75%), nilai minimum, dan maksimum. Tabel 1.7 merupakan rangkuman dari hasil visualisasi pada Gambar 1.7. dalam bentuk tabular.

Tabel 1. 7 Tabel Statistik Deskriptif dari Variabel Dependen *DM, ML, MC,* dan *MR*.

Variabel	**Standar Deviasi**	***Mean***	**Kuartil pertama**	**Kuartil kedua / Median**	**Kuartil ketiga**	**Minimum**	**Maksimum**
DM	7.04	5.62	1	1	12	1	20
ML	5.93	5.12	1	1	11	1	20
MC	8.01	7.51	1	1	16	1	20
MR	5.94	5.15	1	1	11	1	20

Dapat dihasilkan asumsi terhadap data deskriptif dari Gambar 1.6 dan Tabel 1.6. Posisi gelandang *ML* (*Midfielder Left*) merupakan posisi yang paling sedikit dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terkecil dibandingkan dengan posisi bertahan lainnya. Di sisi lain, posisi gelandang *MC* (*Midfielder Center*) merupakan posisi yang paling banyak dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terbesar dibandingkan dengan posisi bertahan lainnya. Dari data juga dapat diasumsikan bahwa banyak dari pemain tidak memiliki keahlian sama sekali pada banyak posisi gelandang, dibuktikan dengan banyak data bernilai 1.



Gambar 1. 8 Diagram Batang Statistik Deskriptif dari Variabel Dependen *AML, AMC, AMR,* dan *ST*

Gambar 1.8. merupakan statistik deskriptif dari variabel dependen *AML, AMC, AMR,* dan *ST*. Statistik deskriptif pada Gambar 1. 8 menunjukkan statistik *mean*, standar deviasi, kuartil pertama (25%), kuartil kedua / median (50%), kuartil ketiga (75%), nilai minimum, dan maksimum. Tabel 1.8 merupakan rangkuman dari hasil visualisasi pada Gambar 1.8. dalam bentuk tabular.

Variabel	Standar Deviasi	*Mean*	Kuartil pertama	Kuartil kedua / Median	Kuartil ketiga	Minimum	Maksimum
AML	7.38	6.92	1	1	14	1	20
AMC	7.39	7.07	1	1	14	1	20
AMR	7.32	6.81	1	1	14	1	20
ST	7.53	6.23	1	1	13	1	20

Dapat dihasilkan asumsi terhadap data deskriptif dari Gambar 1.7 dan Tabel 1.7. Posisi menyerang *ST* (*Striker*) merupakan posisi yang paling sedikit dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terkecil dibandingkan dengan posisi bertahan lainnya. Di sisi lain, posisi gelandang *AMC* (*Attacking Midfielder*) merupakan posisi

yang paling banyak dikuasai oleh pemain sepakbola dengan *mean*, kuartil pertama, median, dan kuartil ketiga terbesar dibandingkan dengan posisi bertahan lainnya. Dari data juga dapat diasumsikan bahwa banyak dari pemain tidak memiliki keahlian sama sekali pada banyak posisi gelandang, dibuktikan dengan banyak data bernilai 1.

b. Analisis Multivariat

Analisis multivariat merujuk pada teknik statistik yang melibatkan analisis simultan dari beberapa variabel. Analisis multivariat mempertimbangkan hubungan timbal balik di antara beberapa variabel. Tujuan utama dari analisis multivariat mencakup pemahaman pola, ketergantungan, dan interaksi di antara variabel-variabel. Ini sering digunakan untuk mengeksplorasi hubungan yang kompleks dalam dataset dengan beberapa variabel.

Dalam kasus ini, analisis multivariat dilakukan untuk mempelajari hubungan antara semua variabel independen dengan variabel dependen. Hubungan antar variabel dilakukan dengan uji hipotesis koefisien korelasi Spearman. Uji hipotesis koefisien korelasi Spearman dipilih karena data tidak berdistribusi normal, sehingga tidak memenuhi asumsi dari uji hipotesis koefisien korelasi Pearson. P-value dalam uji hipotesis yang kurang dari 0.05 menandakan bahwa terdapat hubungan antar dua variabel.

Uji hipotesis koefisien korelasi Spearman dilakukan menggunakan bahasa pemrograman Python. Variabel independen berupa keahlian pemain seperti: : "Corners", "Crossing", "Dribbling", "Finishing", "First Touch", "Free Kick Taking", "Heading", "Long Shots", "Passing", "Tackling", "Technique", "Concentration", "Vision", "Decision", "Determination", "Position.1", "Teamwork", "Balance", "Natural Fitness", "Pace", "Stamina", "Strength", “Left Foot”, “Right Foot” akan diujikan dengan variabel dependen yaitu posisi pemain seperti: "DL", "DC", "DR", "WBL", "WBR", "DM", "MR", "ML", "MC", "AML", "AMC", "AMR", "ST." Kode Python untuk menguji hipotesis koefisien korelasi Spearman ditunjukkan pada Gambar 1.9.

```python
def top_skills_by_position(df, positions, skills):

    position_skills_dict = {}

    for position in positions:
        spearman_corr, p_values = [], []

        for skill in skills:
            spearman_corr_val, p_value = scipy.stats.spearmanr(df[skill], df[position])
            spearman_corr.append(spearman_corr_val)
            p_values.append(p_value)

        results_df = pd.DataFrame({'Skill': skills, 'Spearman_Corr': spearman_corr, 'P_Value': p_values})
        significant_skills = results_df[results_df['P_Value'] < 0.05].sort_values(
            by='Spearman_Corr', ascending=False).head(top_n)['Skill'].tolist()

        position_skills_dict[position] = significant_skills

    return position_skills_dict
```

Gambar 1. 9 Kode Python Uji Hipotesis Koefisien Korelasi Spearman

Output dari uji hipotesis koefisien korelasi Spearman merupakan pasangan keahlian dengan posisi pemain yang memiliki hubungan menurut uji hipotesis. Pasangan keahlian dengan posisi pemain tersebut ditunjukkan pada Gambar 1. 10.

```
{'Tackling': ['DL', 'DC', 'DR', 'WBL', 'WBR', 'DM'],
 'Left Foot': ['DL', 'WBL', 'ML'],
 'Position.1': ['DL', 'DC', 'DR', 'WBL', 'WBR', 'DM', 'MC'],
 'Stamina': ['DL', 'DC', 'DR', 'WBL', 'WBR', 'DM'],
 'Crossing': ['DL', 'WBL', 'WBR', 'MR', 'ML', 'AML', 'AMC', 'AMR'],
 'Teamwork': ['DL', 'DC', 'DR', 'WBL', 'WBR', 'DM', 'MC'],
 'Pace': ['DL', 'DR', 'WBL', 'WBR', 'MR', 'ML', 'AML', 'AMR', 'ST'],
 'Concentration': ['DL', 'DC', 'DR', 'WBR', 'DM'],
 'Strength': ['DL', 'DC', 'DR', 'DM'],
 'Natural Fitness': ['DL', 'DC', 'DR', 'WBL', 'WBR'],
 'Heading': ['DC', 'DR', 'ST'],
 'Balance': ['DC'],
 'Determination': ['DC', 'WBL', 'WBR'],
 'Right Foot': ['DR', 'WBR', 'DM', 'ST'],
 'Corners': ['WBL', 'MR', 'ML', 'MC', 'AML', 'AMC', 'AMR', 'ST'],
 'Passing': ['DM', 'MC', 'AMC'],
 'Decision': ['DM', 'MC'],
 'Vision': ['DM', 'MR', 'ML', 'MC', 'AML', 'AMC', 'AMR'],
 'Dribbling': ['MR', 'ML', 'AML', 'AMC', 'AMR', 'ST'],
 'Technique': ['MR', 'ML', 'MC', 'AML', 'AMC', 'AMR', 'ST'],
 'Free Kick Taking': ['MR', 'ML', 'MC', 'AML', 'AMC', 'AMR', 'ST'],
 'Long Shots': ['MR', 'ML', 'MC', 'AML', 'AMC', 'AMR', 'ST'],
 'Finishing': ['MR', 'AML', 'AMC', 'AMR', 'ST'],
 'First Touch': ['MR', 'ML', 'MC', 'AML', 'AMC', 'AMR', 'ST']}
```

Gambar 1. 10 Pasangan Antara Keahlian dengan Posisi Pemain yang Memiliki Hubungan Menurut Uji Hipotesis Koefisien Korelasi Spearman

Dari hasil gambar 1. 10, variabel dependen yaitu keahlian pemain dapat dikelompokkan menjadi dua jenis, yaitu keahlian bertahan dan keahlian menyerang. Pembagian keahlian bertahan dan menyerang dapat dilihat pada Tabel 1. 8. Tujuan dari pembagian variabel pada kategori bertahan dan menyerang adalah untuk mengurangi dimensi variabel independen,

dikarenakan “kutukan dimensi” dalam *data mining*. Jumlah dimensi yang lebih tinggi secara teoretis memungkinkan penyimpanan informasi yang lebih banyak, tetapi pada praktiknya jarang membantu karena kemungkinan kebisingan (noise) dan redundansi yang lebih tinggi dalam data dunia nyata (Venkat, 2018).

Tabel 1. 8 Pembagian Variabel Bertahan dan Menyerang

Jenis Variabel	**Variabel**
Bertahan	Tackling, Positioning, Stamina, Teamwork, Natural Fitness
Menyerang	Corners, Dribbling, Technique, Free Kick Taking, Long Shots, Finishing, First Touch

2. Data Preparation

Fase persiapan data mencakup semua kegiatan untuk membangun dataset final (data yang akan dimasukkan ke dalam alat pemodelan) dari data mentah awal. Tugas persiapan data kemungkinan akan dilakukan beberapa kali dan tidak dalam urutan tertentu. Tugas-tugas tersebut melibatkan pemilihan tabel, rekaman, dan atribut, pembersihan data, pembangunan atribut baru, dan transformasi data untuk alat pemodelan.

a. Mengatasi Variabel Rendundan

Variabel redundan diatasi dengan mengabaikan atau menggabungkan variabel. Variabel *Strength* akan diabaikan dikarenakan redundan dengan variabel *Balance*. Variabel *Passing, Vision*, dan *Corners* digabungkan menjadi variabel *Playmaking* dikarenakan variabel tersebut dalam konteks sepakbola dapat direpresentasikan menjadi keahlian *Playmaking*. Mengatasi variabel redundan ini bertujuan untuk mengurangi dimensi untuk menghindari “kutukan dimensi” seperti yang sudah disebutkan sebelumnya.

b. Pengelompokan Variabel Dependen

Variabel dependen yaitu posisi pemain akan dikelompokkan menjadi 3 kategori, yaitu bertahan, gelandang, dan penyerang. Pengelompokan ini dilakukan berdasarkan atas konteks sepakbola dimana posisi bermain dibagi menjadi 3 bagian di lapangan, yaitu bertahan, gelandang, dan penyerang. Pengelompokan ini ditujukan untuk meningkatkan performa dari model *machine learning* yang akan dibuat. Pengelompokan posisi pemain menjadi 3 kategori ini ditunjukkan pada Tabel 2. 1.

Tabel 2. 1 Pengelompokan Posisi Pemain

Kelompok	Posisi Pemain
Bertahan	'DL', 'DC', 'DR', 'WBL', 'WBR'
Gelandang	'DM', 'MR', 'ML', 'MC'
Penyerang	'AML', 'AMC', 'AMR', 'ST'

c. Pembuatan Variabel Target Baru

Dibuat variabel dependen baru berdasarkan pengelompokan variabel dependen yang sudah dijelaskan pada bagian b, bab 2. Variabel dependen baru ini merupakan kelompok posisi pemain,"posisi_bertahan," "posisi_gelandang," dan "posisi_penyerang." Variabel dependen ini akan direpresentasikan dengan *one-hot encode* dimana pemain yang masuk dalam kelompok tersebut akan direpresentasikan oleh angka 1 dan pemain yang tidak masuk direpresentasikan oleh angka 0. Pemain dinyatakan masuk dalam sebuah kelompok jika pemain tersebut memiliki ukuran keahlian lebih dari 12 pada salah satu posisi yang bersangkutan. Sebagai contoh, pemain yang memiliki posisi *DC* dengan keahlian 14 akan masuk dalam kelompok "Posisi_Bertahan." Pemain dapat masuk ke dalam lebih dari 1 kelompok jika memang memenuhi syarat. Contoh tabel variabel target baru ditunjukkan pada Gambar 2. 1.

	posisi_bertahan	posisi_gelandang	posisi_menyerang
0	0	1	1
1	0	0	1
2	0	0	1
3	0	0	1
4	0	0	1
...	...	...	...
8447	1	0	0
8448	0	0	1
8449	1	0	0
8450	0	0	1
8451	0	0	1

Gambar 2. 1 Variabel Dependen Baru

d. Pemilihan Variabel Independen dan Dependen

Variabel dependen dikelompokkan menjadi 3 jenis, yaitu variabel dependen kelompok posisi pemain, variabel dependen kelompok posisi bertahan, variabel dependen posisi gelandang, dan variabel dependen penyerang. Pengelompokan ini ditunjukkan pada Tabel 2.2.

Tabel 2. 2 Pengelompokan Variabel Dependen

Kelompok	Variabel Dependen
Posisi Pemain	“posisi_bertahan,” “posisi_gelandang,” dan “posisi_penyerang.”
Bertahan	'DL', 'DC', 'DR', 'WBL', 'WBR'
Gelandang	'DM', 'MR', 'ML', 'MC'
Penyerang	'AML', 'AMC', 'AMR', 'ST'

Variabel independen dibagi menjadi 2 kelompok, yaitu variabel independen untuk variabel dependen kelompok posisi pemain, seperti “posisi_bertahan,” “posisi_gelandang,” dan “posisi_penyerang.” dan variabel independen untuk variabel dependen posisi pemain tanpa pengelompokan seperti, 'DL', 'DC', 'DR', 'WBL', 'WBR', 'DM', 'MR', 'ML', 'MC', 'AML', 'AMC', 'AMR', 'ST'. Pengelompokan ini ditunjukkan pada Tabel 2.3.

Tabel 2. 3 Pengelompokan Variabel Independen

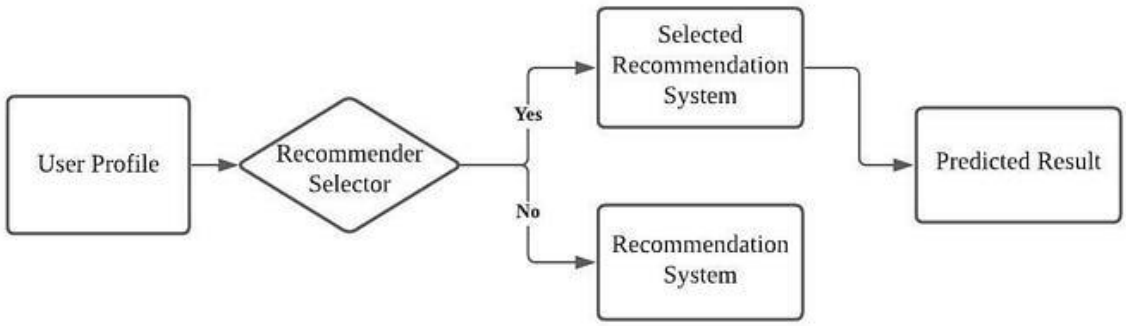
Kelompok	Variabel Dependen
Kelompok Posisi Pemain	“bertahan,” “menyerang”
Tanpa Kelompok Posisi Pemain	'Left Foot', 'Balance', 'Vision', 'Passing', 'Right Foot', 'Concentration', 'Corners', 'Heading', 'Pace', 'Determination', 'Decision', 'Strength'

3. Modelling

Dalam fase *modelling*, berbagai teknik pemodelan dipilih dan diterapkan, serta parameter-parameter mereka dikalibrasi ke nilai-nilai optimal. Terdapat beberapa teknik untuk jenis masalah *data mining* yang serupa. Sebagai contoh, *linear regression* untuk asumsi data berdistribusi normal dan *tree-based model* untuk asumsi data berdistribusi non-normal.

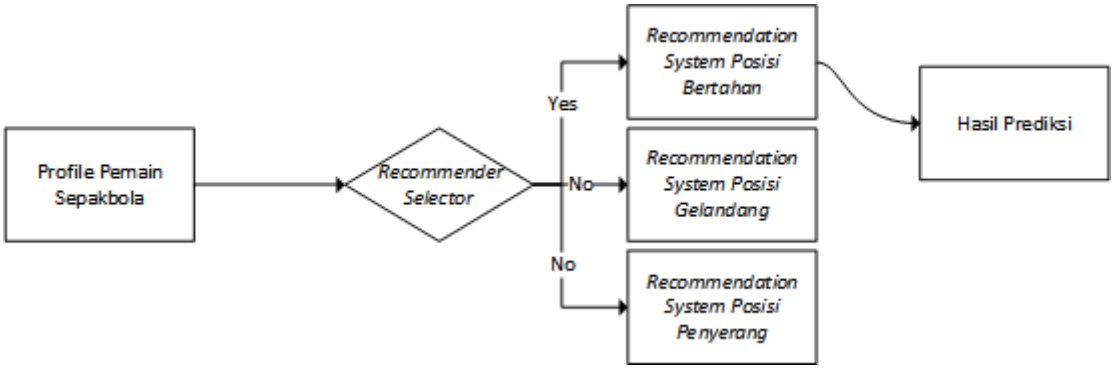
Modelling dilakukan dengan melakukan metode *hybrid recommendation system*. Pendekatan ini bertujuan untuk mengurangi kelemahan dari *Collaborative Filtering* maupun *Content Based* dan mendapatkan keuntungan dari kelebihan keduanya dengan mengintegrasikan dua atau lebih komponen rekomendasi atau implementasi algoritma dalam satu sistem rekomendasi tunggal untuk meningkatkan akurasi *recommendation system* dan mencapai kinerja yang lebih baik.

Metode *hybrid recommendation system* yang dipilih adalah metode *Switching*. Hybrid switching memilih satu sistem rekomendasi berdasarkan kondisi tertentu. Model dibangun untuk dataset yang sensitif terhadap tingkat item, kriteria pemilih rekomendasi dipilih berdasarkan profil pengguna atau fitur lainnya. Pendekatan hybrid switching memperkenalkan lapisan tambahan di atas model rekomendasi, yang memilih model yang sesuai untuk digunakan. Ilustrasi model *hybrid recommendation system* adalah seperti pada Gambar 3.1.



Gambar 3. 1 Metode *Switching* pada *Hybrid Recommendation System*.

Dibuatlah *Switching Hybrid Recommendation System* untuk dataset pada kasus ini dengan 2 jenis model, yaitu K-NN Classifier dan XGBoostRegressor. K-NN Classifier dipilih menjadi *Recommender Selector* dan XGBoostRegressor dipilih menjadi model *recommendation system*. *Recommender Selector* akan memilih *Recommendation System* yang sesuai berdasarkan pembagian posisi pemain seperti yang sudah dijelaskan pada bagian c, subbab 2, sehingga output dari *Recommender Selector* adalah, “posisi_bertahan,” “posisi_gelandang,” dan atau “posisi_penyerang.” Terdapat 3 *recommendation system* yang dapat menjadi pilihan berdasarkan *Recommender Selector*, yaitu *recommendation system* untuk posisi bertahan berupa, 'DL', 'DC', 'DR', 'WBL', 'WBR', lalu untuk posisi gelandang berupa 'DM', 'MR', 'ML', 'MC', dan posisi penyerang berupa 'AML', 'AMC', 'AMR', 'ST'. Diagram flowchart dari metode *Switching Hybrid Recommendation System* terlihat pada Gambar 3.2.



Gambar 3. 2 Flowchart Switching Hybrid Recommendation System yang Dipakai.

4. Evaluation

Evaluasi pertama dilakukan untuk model *Recommender Selector* yaitu K-NN Classifier. *Recommender Selector* mengeluarkan 3 output yaitu label pada "posisi_bertahan," "posisi_gelandang," dan "posisi_penyerang." Evaluasi pertama ini dilakukan dengan *metrics* berupa *hamming loss*. *Hamming loss*, dalam konteks klasifikasi multilabel, adalah metrik evaluasi yang mengukur seberapa banyak elemen yang dihasilkan oleh model klasifikasi tidak sesuai dengan label yang seharusnya. Metrik ini diukur sebagai fraksi dari label yang tidak sesuai dengan label sebenarnya terhadap jumlah total label. Dengan kata lain, Hamming loss memberikan informasi tentang sejauh mana model mengalami kesalahan dalam memprediksi label-label yang seharusnya ada untuk setiap sampel data. Semakin rendah nilai Hamming loss, semakin baik performa model. Hasil evaluasi pertama ditunjukkan pada Tabel 4.1.

Tabel 4. 1 *Hamming Loss Recommender Selector*

Recommender Selector	***Hamming Loss***
K-NN Classifier	0.230

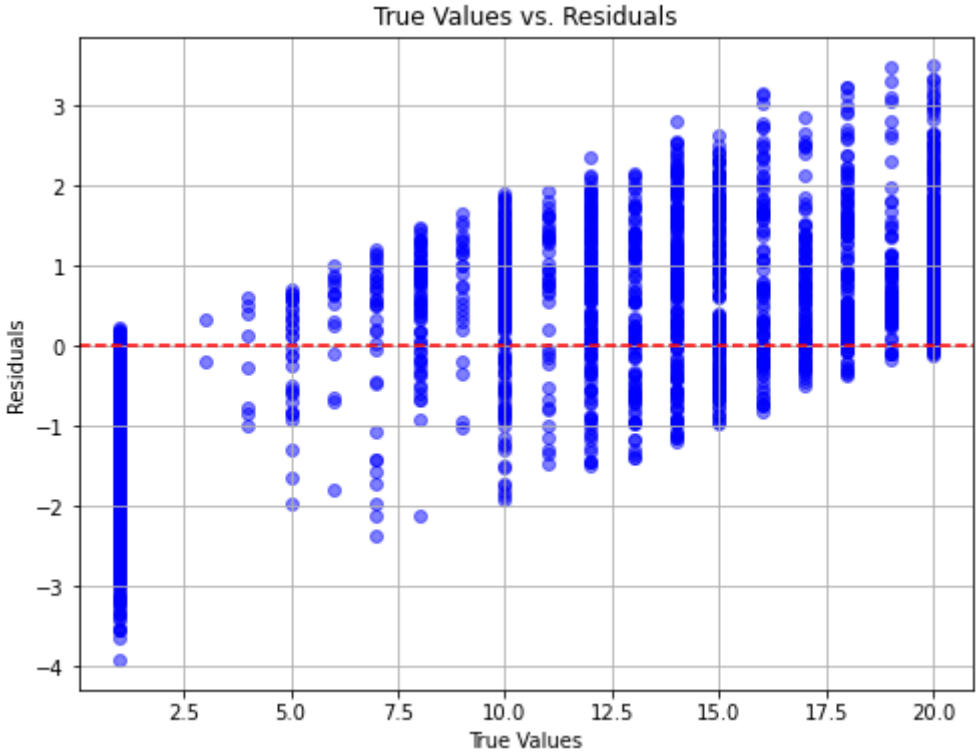
Evaluasi selanjutnya dilakukan untuk model *Recommender System* yaitu XGBoostRegressor. *Recommender System* mengeluarkan output yaitu besaran keahlian pada posisi sepakbola sesuai dengan kelompok posisinya. Sebagai contoh *Recommender System* Posisi Bertahan akan memiliki keluaran berupa besaran keahlian dari posisi 'DL', 'DC', 'DR', 'WBL', 'WBR'. Evaluasi ini akan dilakukan dengan *metrics* berupa *Mean Absolute Error* (MAE). *Mean Absolute Error* (MAE) adalah metrik evaluasi yang digunakan dalam statistika dan machine learning untuk mengukur seberapa dekat prediksi suatu model dengan nilai sebenarnya. MAE dihitung dengan mengambil selisih absolut antara setiap prediksi dan nilai

sebenarnya, lalu mengambil rata-rata dari seluruh selisih tersebut. Hasil evaluasi model *Recommendation System* XGBoostRegressor ditunjukkan pada Tabel 4.2.

Tabel 4. 2 MAE Model *Recommendation System*

Recommendation System	***Mean Absolute Error* (MAE)**
XGBoostRegressor Posisi Bertahan	3.73
XGBoostRegressor Posisi Gelandang	5.02
XGBoostRegressor Posisi Penyerang	5.1

Digunakan analisis residu pada evaluasi terakhir. Analisis residu dalam konteks statistika atau analisis regresi mengacu pada evaluasi dan pemeriksaan sisa atau kesalahan prediksi yang dihasilkan oleh model terhadap data observasional. Residu adalah perbedaan antara nilai yang diprediksi oleh model dan nilai yang sebenarnya dari data yang diamati. Analisis residu bertujuan untuk memastikan bahwa model regresi atau statistika sesuai dengan data dengan baik. Hasil analisis residu terlihat pada Gambar 4.1.



Gambar 4. 1 Hasil Prediksi dengan Residu

Berdasarkan hasil evaluasi yang sudah dipaparkan di atas terdapat beberapa kesimpulan yang dapat diambil. Kesimpulan pertama adalah *recommender selector* berfungsi dengan cukup baik dengan *hamming loss* kurang dari 0.33 atau model tidak memiliki kesempatan

100% dalam kesalahan melabeli salah satu dari 3 kemungkinan label. Kesimpulan kedua adalah performa MAE pada model *recommendation system* posisi bertahan dapat diterima karena memiliki MAE yang cukup, yaitu 3.73. Kesimpulan ketiga adalah performa MAE pada model *recommendation system* posisi gelandang dan penyerang tidak dapat diterima karena memiliki MAE yang buruk, yaitu 5.02 dan 5.1 secara berurutan. Kesimpulan keempat adalah analisis residu menunjukan pola heteroskedastis atau variabilitas tidak konstan, sehingga model tidak dapat dipercaya atau *unreliable* dalam memberikan hasil prediksi. Pola heteroskedastis atau heteroskedastisitas merujuk pada variabilitas yang tidak konstan dari residu atau kesalahan prediksi dalam analisis regresi atau statistika. Dalam hal ini, varians dari residu berubah-ubah sepanjang rentang nilai prediksi. Secara visual, hal ini dapat terlihat sebagai pola yang tidak seragam atau perubahan dalam sebaran residu ketika nilainya berkisar dari rendah ke tinggi atau sebaliknya. Kesimpulan kelima adalah diperlukan pengulangan proses *data mining* untuk menemukan solusi baru atau mencari data lain yang serupa untuk menghasilkan hasil prediksi