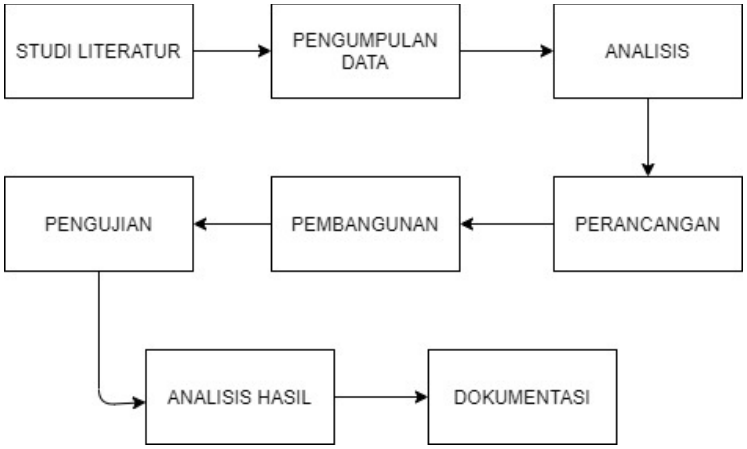
BAB III
METODOLOGI PENELITIAN

Pada bab ini akan dijelaskan mengenai tahapan penelitian yang akan dilakukan. Mulai dari desain penelitian, metode penelitian, maupun alat dan bahan penelitian.

3.1 Desain Penelitian

Berdasarkan desain penelitian pada Gambar 3.1, dapat dilihat bahwa dalam melakukan penelitian ini, ada 8 tahapan yang dilalui. Tahapan-tahapan tersebut adalah studi literatur, pengumpulan data, analisis, perancangan, pembangunan, pengujian, analisis hasil, dan dokumentasi.



Gambar 3.1 Desain penelitian

3.1.1 Studi Literaturs

Studi literatur merupakan tahap pertama yang dilakukan dalam penelitian. Pada tahap ini dilakukan pencarian dan pengkajian terhadap literatur yang terkait dengan penelitian. Studi literatur ini berguna untuk mendalami topik yang diteliti dan juga untuk memahami metode-metode yang digunakan.

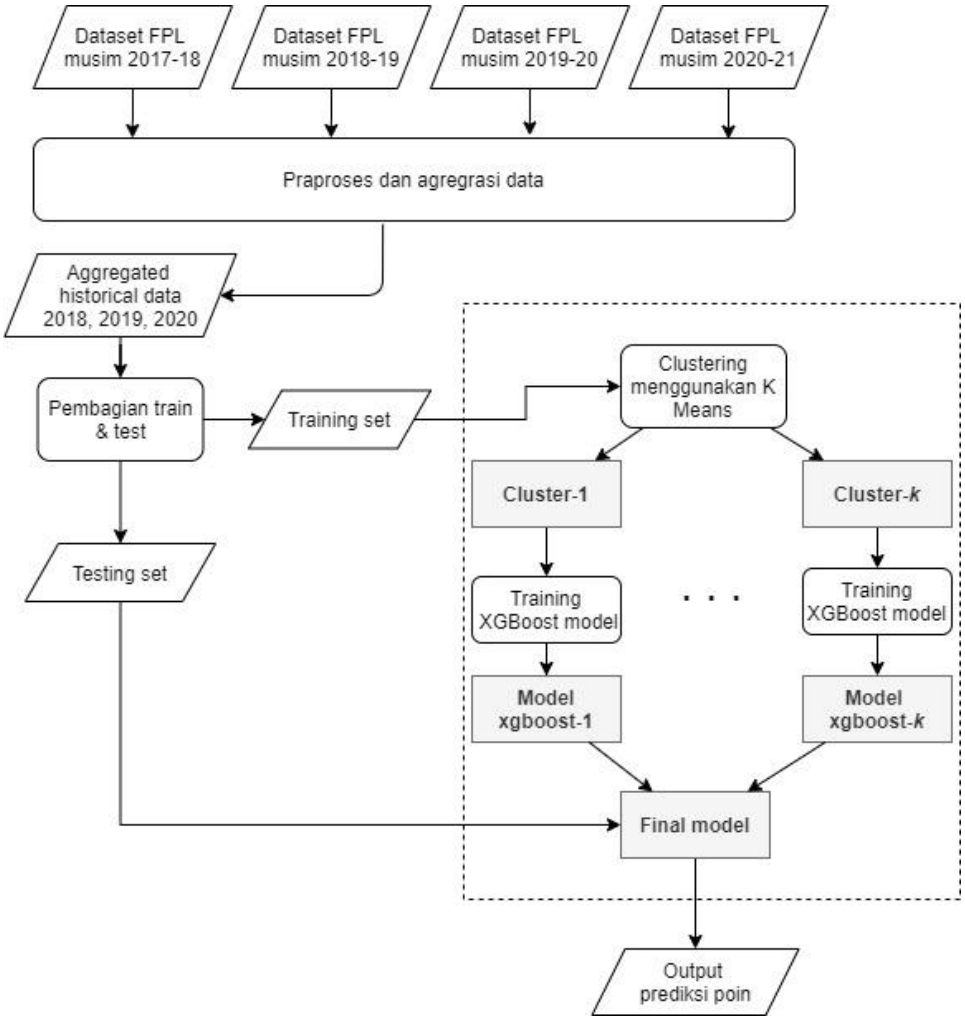
3.1.2 Pengumpulan Data

Tahap pengumpulan data berisi dengan aktivitas pencarian dan pengumpulan data yang dapat digunakan untuk penelitian. Pada penelitian ini, data utama yang digunakan adalah dataset Fantasy Premier League. Website resmi Fantasy Premier League sendiri menyediakan API yang dapat diakses oleh publik

untuk mendapatkan data dari *database* mereka. Namun, website Fantasy Premier League hanya menyediakan data untuk musim yang sedang berlangsung dan tidak menyediakan data untuk musim yang telah berlalu. Pada penelitian ini penulis menggunakan dataset Fantasy Premier League yang diambil dari *repository* Github (Anand, 2021). *Repository* tersebut menyimpan data Fantasy Premier League untuk musim-musim sebelumnya.

3.1.3 Analisis

Dalam tahap ini, data yang dikumpulkan dianalisis terlebih dahulu sebelum digunakan. Analisis data ini dilakukan untuk mengetahui kelengkapan serta karakteristik data sehingga dapat memudahkan tahap selanjutnya.



Gambar 3.2 Diagram rancangan implementasi

3.1.4 Perancangan

Tahap perancangan di sini merupakan tahap yang dilakukan sebelum memulai membangun model. Pada tahap ini peneliti merancang sistem yang akan dibangun secara keseluruhan. Gambara desain rancangan implementasi sistem dapat dilihat pada Gambar 3.2.

3.1.5 Pembangunan

Pada tahap ini, sistem prediksi dibangun dengan mengikuti rancangan yang dibuat pada tahap sebelumnya. Tahap ini dimulai dari pra proses dari data yang dikumpulkan sehingga pada proses pemodelan menggunakan model XGBoost. Penjelasan dari setiap aspek rencana implementasi pada Gambar 3.2 dijelaskan sebagai berikut:

1. Input Data

Data yang digunakan pada penilitian ini adalah data Fantasy Premier League dalam empat musim, yaitu musim 2017, 2018, 2019 dan 2020. Musim sebelum 2017 tidak digunakan karena data di musim-musim tersebut banyak yang tidak lengkap. Seperti yang dijelaskan pada subbab 3.1.2, dataset diambil dari repository Github milik Vaastav Anand (Anand, 2021). Terdapat dua buah jenis data yang digunakan, yaitu data pertandingan Liga Premier Inggris serta data statistik pemain di setiap pertandingan. Jenis data pertama berada di dua buah file yang bernama *teams.csv* dan *fixtures.csv*, sedangkan jenis data kedua berada di file bernama *merged_gw.csv*. Total terdapat tiga buah file yang digunakan pada penelitian. Data yang terdapat pada tiga file tersebut berisi fitur-fitur utama dari data Fantasy Premier League.

Dalam file *team.csv* terdapat 20 baris data, sesuai dengan jumlah tim yang berada di Liga Premier Inggris. Sedangkan untuk *fixtures.csv*, data berjumlah 380 baris karena 20 tim yang ada bertanding sebanyak 38 kali (Andrews, 2019). Nilai ukuran kekuatan yang ada pada tim Liga Premier Inggris ini diberikan oleh pihak Fantasy Premier League menggunakan metode perhitungan yang mereka miliki. Data pertandingan diambil dari dua file tersebut karena pada data di file fixture.csv terdapat fitur id tim yang merujuk ke data di file *teams.csv*.

Detail deskripsi fitur-fitur yang ada pada data *teams.csv* dan *fixture.csv* dapat dilihat masing-masing pada Tabel 3.1 dan Tabel 3.2. Data yang terdapat di file *merged_gw.csv* berisi data statistik pemain Liga Premier Inggris di setiap pertandingan yang telah berjalan. Untuk detailnya, penjelasan fitur pada data tersebut dapat dilihat di Tabel 3.3.

Tabel 3.1
Fitur data *team.csv*

Kolom	*Deskripsi*
id	ID tim
name	Nama tim
strength	Indeks kekuatan dari tim (1 sampai 5)
strength_overall_home	Kekuatan overall tim saat bertanding di kandang sendiri
strength_overall_away	Kekuatan overall tim saat bertanding di kandang lawan
strength_attack_home	Kekuatan serangan tim saat bertanding di kandang sendiri
strength_attack_away	Kekuatan serangan tim saat bertanding di kandang lawan
strength_defence_home	Kekuatan pertahanan tim saat bertanding di kandang sendiri
strength_defence_away	Kekuatan pertahanan tim saat bertanding di kandang lawan

Tabel 3.2
Fitur data *fixtures.csv*

Kolom	*Deskripsi*
id	ID pertandingan
event	Nomor *gameweek* (pekan pertandingan)
team_h	ID tim kandang (home) yang merujuk ke *team.csv*
team_a	ID tim tandang (away) yang merujuk ke *team.csv*
team_h_difficulty	Kekuatan overall tim kandang
team_a_difficulty	Kekuatan overall tim tandang
team_h_score	Skor untuk tim kandang
team_a_score	Skor untuk tim tandang

Tabel 3.3
Fitur data *merged_gw.csv*

Kolom	*Deskripsi*
first_name	Nama depan pemain
last_name	Nama belakang pemain
team	ID tim pemain
fixture	ID pertandingan yang merujuk ke ID di *fixture.csv*
round	Pekan pertandingan ke-
minutes	Jumlah menit bermain
points	Perolehan poin
bonus	Perolehan poin bonus
bps	Penilai *bonus point system*
goals_scored	Jumlah gol yang dicetak
assists	Jumlah asssist
clean_sheet	Status clean sheet
goals_conceded	Jumlah gol kebobolan
own_goals	Jumlah gol bunuh diri
penalties_missed	Jumlah tendangan penalti yang gagal
penalties_saved	Jumlah tendangan penalti yang diselamatkan (untuk penjaga gawang)
saves	Jumah penyelamatan (untuk penjaga gawang)
yellow_card	Kartu kuning
red_card	Kartu merah
selected	Jumlah manajer FPL yang memilih pemain pada tim *fantasy*-nya
price	Harga pemain
creativity	Nilai kreativitas pemain
influence	Nilai pengaruh pemain terhadap kekuatan tim
threat	Nilai ancaman pemain tersebut ke tim lawan
ict_index	Nilai ICT Index. ICT Index merupakan nilai yang diberikan oleh FPL kepada pemain sebagai bentuk penilain pemain berdasarkan statistik pemain tersebut

2. Praproses dan Agregasi Data

Praproses dilakukan pada set data Fantasy Premier League sebelum data digunakan sebagai data *training* dan data *testing*. Tahapan praproses yang dilakukan adalah pembersihan data dengan melakukan pemeriksaan data yang kosong dan tipe data yang tidak sesuai, serta pembuangan fitur yang tidak digunakan. Kemudian setelah data menjadi bersih, data yang berada di file *teams.csv*, *fixtures.csv*, dan *merged_gw.csv* digabungkan menjadi satu dataset. Karena data Fantasy Premier League dapat dikategorikan sebagai data time series, maka untuk setiap musimnya indeks data akan diundur satu baris dan nama fitur diubah dengan adanya tambahan imbuhan kata *prev_* di awal nama fitur.

Pada data Fantasy Premier League dilakukan proses agregasi sebagai bentuk ekstraksi fitur baru data. Fitur baru yang dihasilkan dari proses agregasi diharapkan mampu meningkatkan kinerja model. Praproses yang dilakukan melibatkan proses agregasi dengan musim sebelumnya. Hal ini menyebabkan dataset akhir berisi data musim 2018, 2019, dan 2020. Proses agregasi menyebabkan data di musim 2017 tidak digunakan. Keluaran dari tahap praproses adalah satu dataset utuh yang sudah siap diberikan kepada model.

4. Pembagian data *training* dan *testing*

Tahap pembagian data *training* dan *testing* dilakukan setelah data selesai diagregasi. Data yang dihasilkan akan dipisah menjadi data *training* untuk melatih model dan data *testing* validasi pengujian model. Pembagian data dilakukan dengan rasion 75:25. Training set yang dihasilkan akan diproses lebih lanjut ke dalam proses clustering. Sedangkan testing set akan disimpan dan baru akan digunakan setelah model selesai dibangun.

6. K-means Clustering dan Training model XGBoost

Proses clustering dilakukan pada himpunan data *training* yang telah dibagi pada tahap sebelumya. Lalu pada tiap cluster yang dihasilkan dari proses clustering sebelumnya, sebuah model XGBoost dilatih dengan menggunakan data yang dikelompokkan ke dalam cluster tersebut. Artinya, setiap cluster akan menghasilkan sebuah model XGBoost. Karena fitur pada data yang digunakan

untuk melatih model di setiap cluster sama, maka pada dasarnya model Karena model dilatih menggunakan data yang memiliki fitur yang sama, pada dasarnya model yang dihasilkan di setiap cluster itu sama namun hanya berbeda pada data yang digunakan untuk melatih model.

7. Penggabungan model XGBoost

Proses sebelumnya menghasilkan model XGBoost sejumlah *k* buah. Model yang sudah dilatih dengan data di masing-masing cluster dapat diuji langsung menggunakan testing set yang dihasilkan pada tahap pembagian training set dan testing set sebelumnya. Namun, setiap model akan menghasilkan prediksi yang bias karena model dilatih dengan kelompok data yang berada di cluster saja. Seperti yang dilakukan oleh (Trivedi dkk, 2011), testing set akan diberikan pada setiap model yang dihasilkan kemudian nilai rata-rata dari hasil prediksi masin-masing model akan menjadi nilai predksi akhir.

8. Testing dan Output Model

Pada tahapan testing, model final hasil penggabungan akan diuji dengan testing set yang sudah dibuat pada tahap pembagian dataset. Karena tidak ada perubahan kolom pada training set, maka date testing set tidak perlu diproses lebih lanjut. Untuk menguji model final dilakukan cukup dengan memberikan testing set pada model. Output dari proses implementasi ini adalah prediksi poin pemain di pertandingan Liga Premier Inggris.

3.1.6 Eksperimen

Dalam memprediksi poin untuk setiap pertandingan yang dilalui oleh pemain Liga Premier Inggris, diimplementasikan beberapa eksperimen pada model untuk dibandingkan hasil kinerjanya. Terdapat tiga buah skenario eksperimen yang dilakukan pada penelitian ini. Ketiga skenario tersebut adalah:

1. Implementasi model *ensemble* K-Means dan XGBoost.

   Skenario implementasi model *ensemble* K-Means dan XGBoost merupakan eksperimen inti pada penelitian ini. Pada eksperimen, model *ensemble* dilatih menggunakan dataset hasil praproses. Langkah yang terdapat di eksperimen mengikuti alur yang telah dijelaskan pada subbab 3.1.5. Dataset yang dihasilkan dari tahap praproses dibagi menjadi data latih dan data uji dengan

rasio 72:25. Data latih lalu digunakan untuk proses *clustering* menggunakan K-Means Clustering. Kemudian untuk setiap *cluster* yang dihasilkan, dibuatkan sebuah model XGBoost yang dilatih menggunakan data yang masuk ke dalam *cluster* tersebut. Untuk menguji prediksi model, data uji diberikan ke model di setiap *cluster* kemudian hasilnya dirata-ratakan. Hasil dari eksperimen menentukan apakah model gabungan mampu menaikkan kinerja atau tidak. Agar dapat dilakukan dengan optimal, eksperimen dilakukan dengan menggunakan rentang nilai *k* 2 sampai 12.

2. Implementasi model *ensemble* K-Means dengan algoritma machine learning lainnya.

   Pada skenario model ensemble K-Means dan algoritma lainnya, eksperimen dilakukan dengan cara yang sama dengan eksperimen *ensemble* K-Means dan XGBoost, namun dengan menggunakan berbagai algoritma *machine learning*. Eksperimen dilakukan dengan tujuan untuk mengetahui apakah pengaruh penggabungan model berbeda dengan yang dihasilkan dari penggabungan K-Means dan XGBoost. Algoritma *machine learning* yang digunakan adalah Random Forest, Decision Tree, dan Gaussian Naïve Bayes.

3. Implementasi model *ensemble* K-Means dan XGBoost menggunakan dataset tanpa fitur tambahan.

   Pada skenario eksperimen terakhir, peneliti melakukan pembangunan model *ensemble* yang sama seperti eksperimen sebelumnya namun dengan menggunakan dataset berbeda. Dataset yang digunakan pada eksperimen sebelumnya memiliki banyak fitur tambahan hasil proses aggregasi. Eksperimen dilakukan untuk mengetahui apakah ekstraksi fitur memiliki pengaruh pada kinerja model. Karena tidak dilakukan proses aggregasi, data Fantasy Premier League musim 2017 ikut digabungkan. Dataset yang digunakan pada eksperimen memiliki jumlah data 91182 baris dengan fitur sebanyak 28 buah.

3.1.7 Pengujian

Pada tahap ini dilakukan pengujian kinerja model yang dikembangkan dengan menghitung nilai *Root Mean Square Error* (RMSE) dan *Mean Absolute Error* (MAE) hasil prediksi data uji.

3.1.8 Analisis Hasil

Pada tahap ini dilakukan analisis hasil dari pengujian sistem. Hasil pengujian model prediksi yang dikembangkan peneliti dibandingkan dengan hasil model prediksi dari penelitian lain. Perbandingan tersebut ditampilkan dalam bentuk tabel sehingga dapat ditarik kesimpulan dari perbedaan hasil yang didapatkan.

3.1.9 Dokumentasi

Tahap terakhir dari penelitian ini adalah dokumentasi, di mana semua hasil yang didapatkan dari penelitian ini didokumentasi dan disusun ke dalam bentuk skripsi.

3.2 Alat dan Bahan Penelitian

Perangkat keras yang digunakan dalam penelitian ini adalah sebuah laptop dengan spesifikasi sebagai berikut:

- Intel i5-4200U CPU @1.60GHz 2.30GHz , dengan 2 core dan 4 thread
- 8GB RAM
- SSD 256GB

Dan perangkat lunak sebagai berikut:

- Sistem Operasi Windows 10 64-bit
- Python 3.7
- Jupyter Notebook