BAB 3
METODOLOGI PENELITIAN

3.1 Telaah Literatur

Pada tahap ini akan mengumpulkan materi atau literatur dari berbagai sumber seperti jurnal, artikel, buku, *conference*, dan sumber lainnya yang berkaitan dengan topik penelitian ini yaitu emisi karbon, *machine learning*, *gradient boosting*, *regression*, *XGBoost Regressor*, listrik.

3.2 Pengumpulan Data

Pada tahap ini akan melakukan pengumpulan data-data dan referensi yang dibutuhkan untuk membentuk *dataset* yang diperlukan untuk melakukan penelitian. *Dataset* yang akan digunakan dalam penelitian ini berasal dari sebuah penelitian terdahulu yang telah dilakukan oleh Luis M. [25] mengenai prediksi konsumsi/penggunaan listrik sebuah rumah yang berada di Belgia. Berikut adalah *detail/informasi mengenai dataset* rumah Belgia tersebut:

- Jumlah *Feature* yang akan digunakan adalah 26 dan *feature* tersusun atas:

Tabel 3.1. Data Variabel dan deskripsi

Feature List	Units
Lights_wh	Wh
dapur_temp	°C
dapur_humid	%
rTamu_temp	°C
rTamu_humid	%
kamar_Tidur_temp	°C
kamar_Tidur_humid	%
kamar_Tidur2_temp	°C
kamar_Tidur2_humid	%
master_bedroom_temp	°C

Tabel 3.2. Data Variabel dan deskripsi (lanjutan)

Feature List	Units
master_bedroom_humid	%
rKerja_temp	°C
rKerja_humid	%
kamar_Mandi_temp	°C
kamar_Mandi_humid	%
luar_Bangunan_temp	°C
luar_Bangunan_humid	%
rSetrika_temp	°C
rSetrika_humid	%
Average_building_Temperature	°C
Average_building_humidity	%
seconds_to_midnight	Seconds
time_weekday	Integer
time_moth	Integer
time_day	Integer
time_hour	Integer

Pemilihan penggunaan energi listrik lampu (Lights_wh) sebagai *feature* dilakukan karena menurut Luis M., lampu merupakan metrik yang telah terbukti baik dalam memprediksi okupansi sebuah ruangan digabung dengan perhitungan kelembapan suatu ruangan, yang dimana besar okupansi tersebut dapat mengindikasikan besar atau kecilnya penggunaan peralatan rumah tangga lainnya [39].

- Label yang akan digunakan berupa angka total penggunaan listrik oleh peralatan rumah tangga (tidak termaksud lampu) dalam satuan Wh.
- Jumlah *Sample* yang akan digunakan adalah 19735 dan *sample* berasal dari pencatatan suhu dan kelembapan dengan menggunakan sensor bernama *ZigBee Wireless Sensor Network*, yang dimana sensor ini dari pabrik akan mencatat suhu dan kelembapan rumah setiap 3.3 menit, lalu data tersebut akan dirata-ratakan untuk setiap 10 menit selama 137 hari. Menurut Luis

M. pemilihan waktu 10 menit dilakukan untuk menangkap atau melihat perubahan cepat dari konsumsi energi peralatan rumah tangga [25].

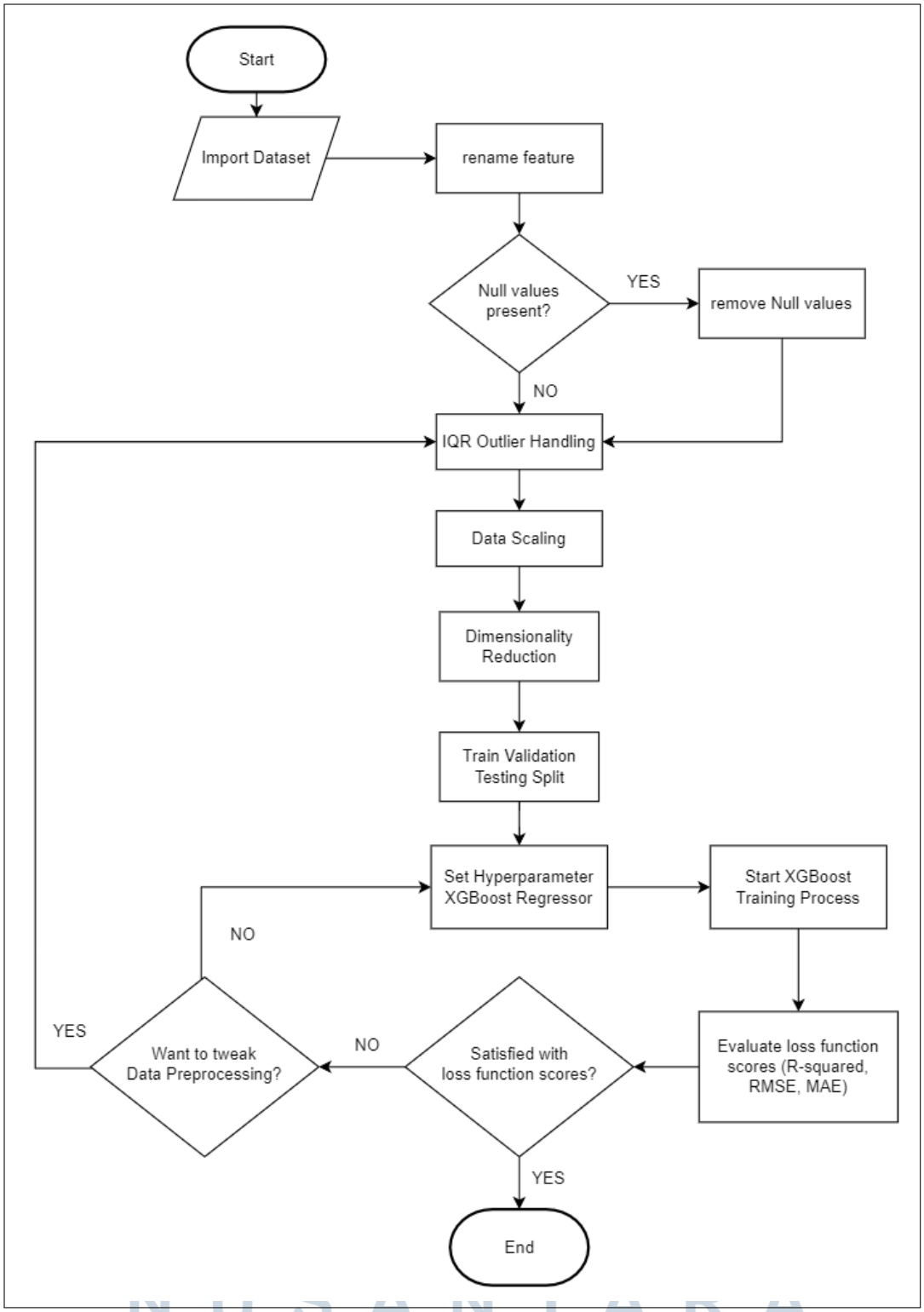
date	appliances_wh	lights_wh	dapur_temp	dapur_humid	rTamu_temp	rTamu_humid	kamar_Tidur_temp	kamar_Tidu
2016-01-11 17:00:00	60	30	19.890000	47.596667	19.200000	44.790000	19.790000	44
2016-01-11 17:10:00	60	30	19.890000	46.693333	19.200000	44.722500	19.790000	44
2016-01-11 17:20:00	50	30	19.890000	46.300000	19.200000	44.626667	19.790000	44
2016-01-11 17:30:00	50	40	19.890000	46.066667	19.200000	44.590000	19.790000	45
2016-01-11 17:40:00	60	40	19.890000	46.333333	19.200000	44.530000	19.790000	45
...	...	...	...	...	...	...	...	
2016-05-27 17:20:00	100	0	25.566667	46.560000	25.890000	42.025714	27.200000	41
2016-05-27 17:30:00	90	0	25.500000	46.500000	25.754000	42.080000	27.133333	41
2016-05-27 17:40:00	270	10	25.500000	46.596667	25.628571	42.768571	27.050000	41
2016-05-27 17:50:00	420	10	25.500000	46.990000	25.414000	43.036000	26.890000	41
2016-05-27 18:00:00	430	10	25.500000	46.600000	25.264286	42.971429	26.823333	41

19735 rows × 22 columns

Gambar 3.1. Data Rumah Belgia

3.3 Gambaran umum proses pembuatan model XGBoost Regressor

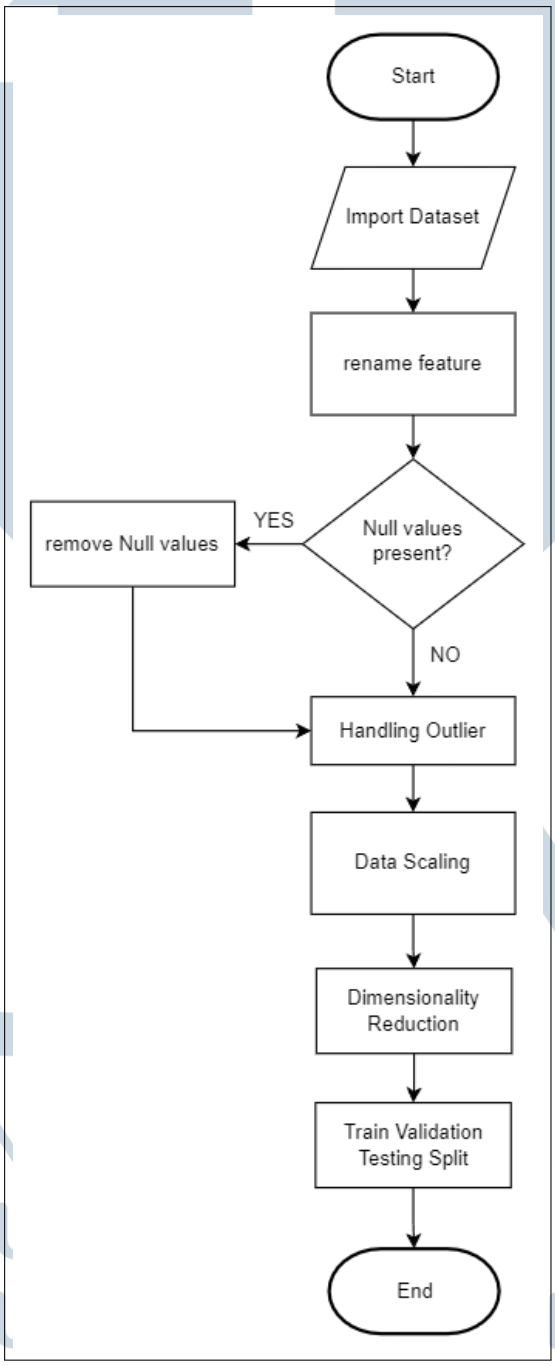
Gambar 3.2 di bawah merupakan *flowchart* proses pembuatan model *XGBoost Regressor*. Proses tersebut dimulai dengan melakukan import dataset, lalu dataset akan melalui proses data preprocessing.



Gambar 3.2. *Flowchart proses pembuatan model XGBoost Regressor*

3.4 Data Preprocessing

Pada tahap ini data-data yang sudah terkumpul harus diolah terlebih dahulu. Berikut adalah *Flowchart Data Preprocessing*.



Gambar 3.3. *Flowchart Data Preprocessing*

3.4.1 Rename Feature

Penamaan ulang *feature* dari nama yang abstrak seperti T1 dan RH_1 menjadi dapur_temp dan dapur_humid dilakukan agar nama feature lebih mudah dibaca dan lebih mudah di identifikasi.

3.4.2 Null Value Checking

Pada tahap ini akan dilakukan proses pengecekan jika ada nilai *null* dalam dataset karena *missing* atau *null value data* dalam *dataset* dapat memengaruhi secara negatif performa *machine learning model* [40].

3.4.3 *Outlier detection and handling*

Proses *Outlier detection and handling* dilakukan dengan tujuan mendeteksi anomali pada *dataset* yang dapat memengaruhi performa dan akurasi model [41]. Metode deteksi *outlier* yang digunakan dalam penelitian ini adalah *Percentile Method*.*Percentile Method* merupakan salah satu metode penanganan *outlier* yang efektif karena memiliki kemampuan yang baik dalam pemisahan *outlier*, dan memiliki performa yang konsisten dalam mendekteksi *outlier* [42].

3.4.4 *Data Scaling*

Proses *Data Scaling* dilakukan agar rentang nilai (*scale*) dalam *dataset* memiliki *scale* yang sama sehingga dapat meminimalisir kesempatan sebuah *feature* mendominasi proses *training* karena skala-nya yang jauh lebih besar dari *feature* lain, yang pada akhirnya akan mengarah pada model yang bias terhadap *feature* yang mendominasi tersebut [43]. Metode *Data Scaling* yang digunakan dalam penelitian ini adalah *MinMax Scaler*

3.4.5 *Dimensionality Reduction*

Proses *Data Scaling* dilakukan dengan tujuan untuk mengurangi ukuran dan kompleksitas *dataset* dengan mempertahankan relasi atau hubungan antar *feature* sehingga dapat meningkatkan performa model *machine learning* [44]. Metode *Dimensionality Reduction* yang digunakan dalam penelitian ini adalah *Principal Component Analysis (PCA)*.

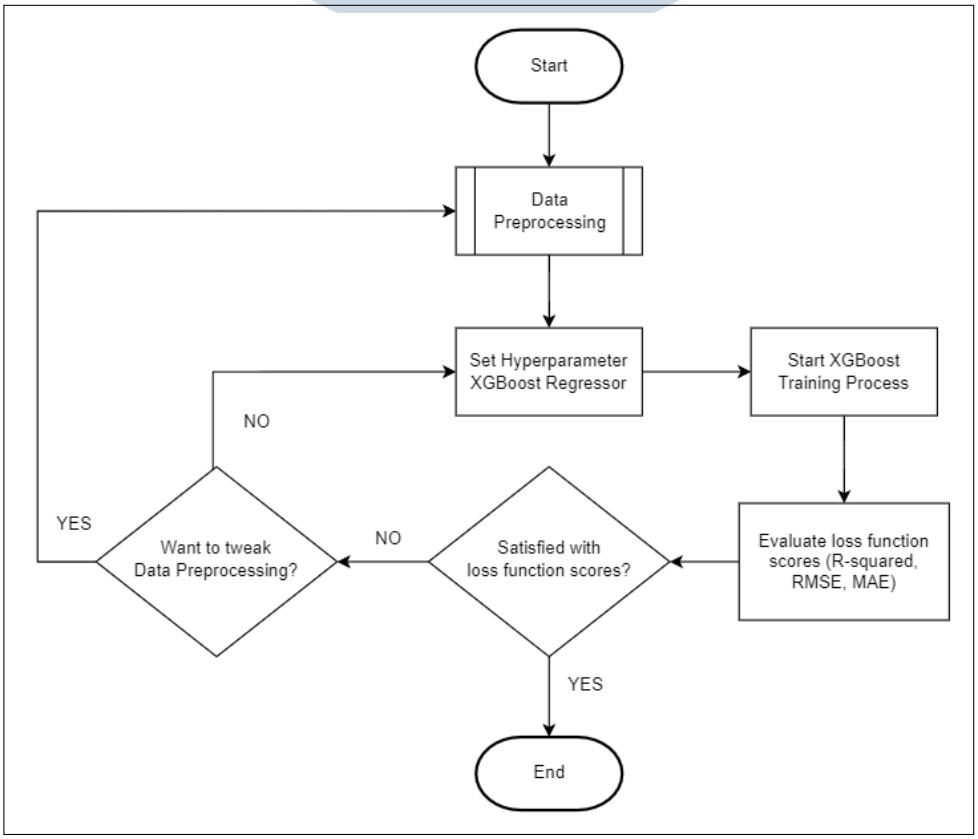
3.4.6 *Test Data*

Pada tahap ini 50 *data sample* dari *dataset* akan diambil secara *random* dan 50 data tersebut akan digunakan untuk mengevaluasi model yang telah selesai training, untuk menilai performa prediksi model ketika diberikan data yang model tersebut belum pernah lihat. Pemilihan jumlah 50 *data sample* dilakukan karena alasan menjadi jumlah dataset untuk *training validation*, sedekat mungkin dengan milik penelitian terdahulu.

3.4.7 *Train Validation Split*

Pada tahap ini dilakukan *train Validation split* terhadap data agar terbagi menjadi dua bagian yaitu *training data* dan *Validation data*, dengan rasio 75:25, yaitu 75% *training* dan 25% *Validation*.

3.5 Training Model



Gambar 3.4. *Flowchart Training Model*

Pada tahap ini perancangan model dimulai dengan memberikan nilai-nilai *hyperparameter* untuk *XGBoost Regressor* model seperti n_estimators berfungsi untuk menentukan batas maksimum proses *boosting* atau jumlah pohon/*tree* yang akan dibentuk oleh model, early_stopping_rounds berfungsi untuk menentukan batas toleransi model terhadap kenaikan nilai loss testing, dengan tujuan untuk menghentikan proses *training* secara prematur jika nilai *loss* mulai meningkat sehingga model dapat terhindar dari *overfitting*, device berfungsi untuk menentukan *hardware* yang akan digunakan untuk proses *training* seperti *CPU* atau *GPU (CUDA)*, dan learning_rate berfungsi untuk menentukan kecepatan *XGBoost Regressor* melakukan perubahan atau *update* terhadap tree yang telah dibuat. Setelah *hyperparameter* telah ditentukan proses *training* atau *fitting XGBoost Regressor* dapat dimulai.

3.6 Uji Coba, Analisis dan Evaluasi

Uji Coba, Analisis dan Evaluasi dilakukan untuk mengetahui peningkatan atau penurunan performa algoritma *XGBoost Regression* dalam memprediksi emisi Karbon, dengan bantuan *loss fucntion* RMSE, MAE, dan *R-Squared*. Uji coba dilakukan dengan membandingkan nilai *loss* sejumlah model yang merupakan hasil dari iterasi *training* yang berbeda dan mengevaluasi perubahan atau *tweaks* yang menghasilkan akurasi terbaik dan nilai *loss* terendah atau terbaik.